

Text Information Retrieval in Tetun

Gabriel de Jesus



Doctoral Program in Informatics Engineering

Supervisor: Prof. Sérgio Sobral Nunes, Ph.D.

September 2025

Text Information Retrieval in Tetun

Gabriel de Jesus

Doctoral Program in Informatics Engineering

Approved in public defense by the following jury:

President: Prof. Rui Filipe Lima Maranhão de Abreu, Ph.D.

External Examiner: Prof. Arjen P. de Vries, Ph.D.

External Examiner: Prof. Bruno Emanuel da Graça Martins, Ph.D.

Internal Examiner: Prof. Henrique Daniel de Avelar Lopes Cardoso, Ph.D.

Supervisor: Prof. Sérgio Sobral Nunes, Ph.D.

September 2025

Abstract

Ensuring access to information in all languages is crucial for bridging disparities in communities' participation in the digital age and fostering a more inclusive and equitable society, particularly for speakers of low-resource languages. However, enabling such access remains a significant challenge for many of these communities. Tetun, a dialect and lingua franca that became one of Timor-Leste's official languages when the country restored its independence in 2002, faces similar challenges. According to the 2015 census, Tetun is spoken by approximately 79% of the country's 1.18 million population. Despite its official status, Tetun remains underserved in language technology. Specifically, information retrieval-based solutions for Tetun do not exist, making it challenging to find relevant information on the internet and digital platforms through text-based search in the language.

This work tackles these challenges by investigating retrieval strategies for text-based search that can enable the application of information retrieval techniques to develop search solutions for Tetun, with a specific focus on the ad-hoc text retrieval task. Given that language-specific algorithms, tools, and document collections for Tetun were previously unavailable, this work began by creating these foundational resources, which serve as contributions relevant to the information retrieval and natural language processing domains. These resources include a tokenizer, a language identification model, a stemmer, a stopword list, a document collection, a test collection, baselines for the ad-hoc text retrieval task, and a search log dataset.

The contributions to information retrieval for low-resource languages include: (1) A data collection pipeline tailored for low-resource languages to streamline the construction of textual data from the web; (2) A human-in-the-loop methodology for annotating, processing, and constructing a dataset well-suited for a variety of information retrieval and natural language processing tasks; (3) A novel network-based approach for stopword detection; (4) Methodologies for developing a stemmer, designed for a language heavily influenced by loanwords, and the construction of a ground truth set for evaluating stemmer performance; (5) A detailed approach for constructing a test collection to evaluate the effectiveness of retrieval systems; (6) A methodology for establishing a robust baseline for the ad-hoc text retrieval task; and (7) Document contextualization and dual-parameter tuning strategies for hybrid text retrieval.

The results from this work contribute to the development of technologies associated with the computational processing of Tetun, address gaps in its linguistic resources, and achieve impactful outcomes that elevate Tetun's status. These advancements open new opportunities for future research and innovation. Moreover, this work introduces promising methodologies that can be adapted to other languages facing similar challenges, thereby contributing to the broader advancement of information retrieval for low-resource languages.

Resumo

Garantir o acesso à informação em todas as línguas é crucial para reduzir as disparidades na participação das comunidades na era digital e promover uma sociedade mais inclusiva e equitativa, em particular para os falantes de línguas com poucos recursos. Contudo, permitir esse acesso continua a ser um desafio significativo para muitas dessas comunidades. O Tetun, um dialeto e língua franca que se tornou uma das línguas oficiais de Timor-Leste aquando da restauração da independência do país em 2002, enfrenta desafios semelhantes. De acordo com o censo de 2015, o Tetun é falado por aproximadamente 79% da população de 1,18 milhões. Apesar do seu estatuto oficial, o Tetun continua a ser subservido na tecnologia linguística. Especificamente, não existem soluções de recuperação de informação para Tetun, o que dificulta acesso a informação relevante na internet e em plataformas digitais através de pesquisa textual nesta língua.

Este trabalho aborda estes desafios, explorando estratégias de recuperação para pesquisas textuais que permitam aplicar técnicas de recuperação de informação no desenvolvimento de soluções de pesquisa para Tetun, com um foco específico na tarefa de recuperação *ad-hoc* de texto. Tendo em conta a inexistência de algoritmos, ferramentas e coleções de documentos para Tetun, este trabalho iniciou-se com a criação desses recursos fundamentais, que constituem contribuições relevantes para os domínios da recuperação de informação e do processamento de linguagem natural. Estes recursos incluem um *tokenizador*, um modelo de identificação de língua, um *stemmer*, uma lista de *stopwords*, uma coleção de documentos textuais em Tetun, uma coleção de julgamentos de relevância, resultados de referência para a tarefa de recuperação *ad-hoc* de texto, e uma coleção de interrogações de pesquisa.

As contribuições para a recuperação de informação em línguas com poucos recursos incluem: (1) Um processo estruturado de recolha de dados, concebido para línguas com poucos recursos, que visa agilizar a construção de dados textuais a partir da web; (2) Uma metodologia com intervenção humana para anotar, processar e construir um conjunto de dados adequado a diversas tarefas de recuperação de informação e processamento de linguagem natural; (3) Uma abordagem baseada em redes para a deteção de *stopwords*; (4) Metodologias para o desenvolvimento de um *stemmer*, concebido para uma língua fortemente influenciada por empréstimos linguísticos, e a criação de um conjunto de dados de referência para a avaliação do seu desempenho; (5) Uma metodologia para a construção de um conjunto de dados de referência para a avaliação dos sistemas de recuperação; (6) Uma metodologia para estabelecer resultados de referência robustos para a tarefa de recuperação *ad-hoc* de texto; e (7) Estratégias de contextualização de documentos e de afinação com dois parâmetros para recuperação de texto baseada em pesquisa híbrida.

Os resultados deste trabalho contribuem para o desenvolvimento de tecnologias associadas ao processamento computacional de Tetun, preenchem lacunas nos seus recursos linguísticos e alcançam resultados que elevam o estatuto do Tetun. Estes avanços abrem novas oportunidades para investigações e inovações futuras. Além disso, este trabalho introduz metodologias adaptáveis a outras línguas que enfrentam desafios semelhantes, contribuindo, assim, para o avanço da recuperação de informação em línguas com poucos recursos.

Rezumu

Garante asesu ba informasaun iha lian hotu-hotu ne'e krusiál atu hamenus disparidade sira ba comunidade sira-nia partisipasaun iha era dijital no promove sosiedade ida-ne'ebé inkluziva no ekitativa, liu-liu ba ema sira-ne'ebé ko'alia lian ho rekursu oituan. Maibé, permite asesu ida-ne'e nafatin sai nu'udar dezafiu signifikativu ida ba comunidade sira-ne'e. Tetun, dialetu no lian *franca* ida-ne'ebé sai Timor-Leste nia lian ofisial ida bainhira nasaun restaura ninia independénsia iha 2002, hasoru mós dezafiu sira-ne'ebé semellante. Haktuir sensu 2015, Tetun nu'udar lian ida-ne'ebé ko'alia husi besik 79% husi populasau hamutuk 1.18 millaun. Maski ho estatutu ofisial, Tetun seidauk servi ho di'ak iha teknolojia linguística. Liu-liu, laiha solusaun bazeia ba rekupersaun informasaun ba Tetun, ida-ne'ebé halo susar atu hetan informasaun relevante iha internet no plataforma dijital sira ba interrogasaun sira ba peskiza ho testu iha lian ida-ne'e.

Servisu ida-ne'e rezolve dezafiu sira-ne'e, investiga estratejia rekupersaun ba peskiza bazeia ba testu ne'ebé bele permite aplikasaun téknika sira rekupersaun informasaun nian atu dezen-volve solusaun peskiza ba Tetun, ho foku espesífiku ba tarefa rekupersaun *ad-hoc* ba testu. Tanba algoritmia, ferramenta, no kolelau dokumentu sira ba Tetun antes ne'e seidauk iha, servisu ida-ne'e komesa ho kriaun rekursu sira-ne'e, ne'ebé serve nu'udar kontibuissau relevante ba iha domínio rekupersaun informasaun no prosesamentu linguajen natural nian. Rekursu sira-ne'e inklui tokenizadór ida, modelu identifikasaun ba lian ida, *stemmer* ida, lista *stopwords* ida, kolelau dokumentu testu ida, *baselines* robusta ba tarefa rekupersaun *ad-hoc* ba testu, no kolelau ida ba interrogasaun peskiza nian.

Kontribuisau sira ba rekupersaun informasaun ba lian ho rekursu oituan inklui: (1) Prosesu estruturadu ba rekolla dadus nian ida-ne'ebé konsebida ba lian ho rekursu oituan, atu hafasil konstrusaun dadus testu husi web; (2) Metodolojia ida ho intervensau umana hodi anota, prosesa, no konstrui dadus ne'ebé adekua ba tarefa sira rekupersaun informasaun no prosesamentu linguajen natural nian; (3) Abordajen foun bazeia ba rede ba detesau *stopwords*; (4) Metodolojia sira hodi dezen-volve *stemmer*, ne'ebé dezenia ba lian sira-ne'ebé influénsia makaas husi em-préstimu, no ba konstrusaun ba konjuntu dadus referénsia atu avalia *stemmer*-nia dezempeñu; (5) Metodolojia ida ba konstrusaun konjuntu dadus referénsia atu avalia sistema rekupersaun sira; (6) Metodolojia ida atu estabelese rezultadu referénsia robustu ba tarefa rekupersaun *ad-hoc* ba testu; no (7) Estratejia ba kontestualizasaun dokumentu no afinasaun ho parámetru rua ba rekupersaun híbrida testu.

Rezultadu hosi servisu ida-ne'e kontribui ba dezen-volvimentu teknolojia asosia ho prosesamentu komputasionál ba Tetun, priense lakuna sira-ne'ebé hasoru iha ninia rekursu linguístiku no alkansa rezultadu impaktante ne'ebé eleva Tetun ninia estatutu. Avansu sira-ne'e loke oportunidade foun ba investigasaun no inovasaun sira iha futuro. Iha sorin seluk, servisu ida-ne'e estabelese metodolojia sira-ne'ebé bele adapta ba lian seluk ne'ebé hasoru problema semellante, nune'e, kontribui ba avansu iha rekupersaun informasaun ba lian sira-ne'ebé ho rekursu oituan.

Acknowledgements

The journey to completing this thesis was marked by both professional challenges and personal struggles. I am deeply grateful to my supervisor, Professor Sérgio Nunes, for his insightful guidance, unwavering support, and patience. His encouragement and expertise were instrumental in overcoming obstacles and refining my work, and his belief in my potential kept me motivated during the most challenging times.

I would also like to express my heartfelt gratitude to my family, particularly Eliseia Seixas, a very special person in my life. My sister, Domingas Amaral, along with the other members of the RIANG family—Matias Costa, Marcos Soares, and Natalina Amaral—have been a constant source of strength and encouragement throughout this journey. I honor the memory of my beloved parents “Ricardo Pinto and Angelina da Costa”, my brother “Silvino Amaral”, and my uncle “Salvador Pinto”, whose presence in my life left an indelible mark. I am equally thankful to Professor Gabriel David and his family for their generous hospitality and kindness, which made me feel at home during the critical phase of my research.

This thesis would not have been possible without the invaluable contributions of many volunteer annotators. I would especially like to acknowledge Fanézio Pinto, Altedio Araújo, Roohy Freitas, Alton Freitas, Rita Belo, and Dominica Lisboa, who represent the broader community of contributors. Their dedication and efforts were essential to the success of this work.

I am deeply grateful for the Ph.D. studentship financed by National Funds through the Portuguese funding agency, Fundação para a Ciência e a Tecnologia (FCT), under the grant number SFRH/BD/151437/2021 (DOI: [10.54499/SFRH/BD/151437/2021](https://doi.org/10.54499/SFRH/BD/151437/2021)), which made this research possible. I also acknowledge the facilities and resources provided by the Institute for Systems and Computer Engineering, Technology and Science (INESC TEC). My special thanks go to Rosa Tilman, Director of the *Instituto Nacional de Linguística* (INL), for her generosity in supporting Tetun-related materials, which served as invaluable resources for this research.

Finally, I extend my sincere gratitude to my colleagues at InfoLab, and to everyone who, in any capacity, contributed to this work and supported me throughout my Ph.D. journey.



Gabriel de Jesus

*“Rising from the shadows of illiteracy and poverty,
I stand as proof that education is a power
capable of rewriting the history of generations”*

*To my sister “Domingas Amaral”
and my son “Richelino Riag”*

Gabriel de Jesus

Contents

1	Introduction	1
1.1	Context and Motivation	2
1.2	Problem Description	3
1.3	Objectives	4
1.4	Research Questions and Hypothesis	5
1.5	Methodology and Key Contributions	6
1.6	Publications and Research Outputs	7
1.6.1	Published Papers	7
1.6.2	Paper Under Review	9
1.6.3	Datasets, Tools, Algorithms, Software, and Prototypes	9
1.7	Thesis Structure	10
2	Background	12
2.1	Information Retrieval	12
2.1.1	Definition	13
2.1.2	Ad-Hoc Text Retrieval	13
2.1.3	Architecture	14
2.1.4	Web Crawling	14
2.1.5	Text Preprocessing	15
2.1.6	Indexing, Retrieval, and Ranking	16
2.1.7	Models	17
2.1.8	Test Collections	23
2.1.9	Evaluation Metrics	24
2.1.10	Notion of Relevance	26
2.2	Tetun	27
2.2.1	Evolution	27
2.2.2	Evolution in Newspapers	28
2.2.3	Orthography	29
2.2.4	Morphology	29
2.2.5	Loanwords	30
2.2.6	Text Documents	31
3	Text Information Retrieval in Low-Resource Languages	32
3.1	Overview	32
3.2	Dataset Construction Pipelines	32
3.2.1	BootCaT Toolkit	33
3.2.2	Tailored Crawling	33
3.2.3	CURL Pipeline	34

3.2.4	Corpulyzer Framework	34
3.2.5	Common Crawl Processing Pipeline	34
3.2.6	SwissCrawl Pipeline	35
3.2.7	Summary	35
3.3	Text Preprocessing Techniques	36
3.3.1	Language Identification	36
3.3.2	Stopword Detection	37
3.3.3	Stemming	37
3.4	Test Collection Development	38
3.5	Ad-hoc Text Retrieval	38
3.6	Summary	39
4	Tetun Text Resource Development	41
4.1	Labadain Crawler Development	41
4.1.1	Tetun Tokenizer	42
4.1.2	Tetun Language Identification	43
4.1.3	Labadain Crawler Architecture	46
4.1.4	Experiment on Constructing a Tetun Text Corpus	48
4.1.5	Experimental Result Analysis	48
4.1.6	Evaluation of the Corpus Quality	49
4.1.7	Discussion	51
4.1.8	Conclusions	52
4.2	Labadain-30k+ Dataset Construction	52
4.2.1	Document Annotation	53
4.2.2	Content and Date Extractions	54
4.2.3	Deduplication and Post Processing	54
4.2.4	Document Characterization	55
4.2.5	Wikipedia Document Processing	56
4.2.6	Final Dataset	57
4.2.7	Discussion	57
4.2.8	Conclusions	58
4.3	Labadain-30k+ Dataset Analysis	59
4.3.1	Dataset Description and Distribution	59
4.3.2	Document Evolution on the Web	61
4.3.3	Analysis on the Results	63
4.3.4	Discussion	64
4.3.5	Conclusions	64
5	Tetun Language Component Building	66
5.1	Stopword Detection	66
5.1.1	Data Collection and Preparation	67
5.1.2	Co-Occurrence Network Properties	67
5.1.3	Traditional Frequency-Based Methods	69
5.1.4	Experiment and Evaluation	70
5.1.5	Results and Discussion	71
5.1.6	Conclusions	74
5.2	Tetun Stopword List Building	74
5.2.1	Methodology	75
5.2.2	Experiment and Evaluation	75

5.2.3	Results	76
5.2.4	Discussion	77
5.2.5	Conclusions	78
5.3	Labadain-Stemmer Development	78
5.3.1	Tetun Affixes	79
5.3.2	Labadain-Stemmer Variants	79
5.3.3	Text Sample Construction for Evaluation	80
5.3.4	Ground Truth Development	82
5.3.5	Evaluation Metric	82
5.3.6	Experimental Setting	83
5.3.7	Evaluation and Results	83
5.3.8	Discussion	85
5.3.9	Limitations	85
5.3.10	Conclusions	85
6	Tetun Test Collection	87
6.1	Overview	87
6.2	Constructing a Test Collection for Tetun	88
6.2.1	Document Collection	88
6.2.2	Query Formulation	89
6.2.3	Document Pooling	90
6.2.4	Relevance Judgments	92
6.2.5	Results and Discussion	96
6.2.6	Conclusions	97
6.3	Exploring LLMs for Relevance Judgments in Tetun	97
6.3.1	Overview	98
6.3.2	Relevance Judgments for Tetun Using LLMs	98
6.3.3	Experiment with Tetun	100
6.3.4	Results and Discussion	102
6.3.5	Conclusions	102
7	Tetun Ad-Hoc Text Retrieval	104
7.1	Preliminary Study	104
7.1.1	Overview	104
7.1.2	Experimental Setup	105
7.1.3	Test Collection Development	105
7.1.4	Evaluation and Results	106
7.1.5	Conclusions	106
7.2	Baselines	107
7.2.1	Overview	107
7.2.2	Text Preprocessing	107
7.2.3	Experimental Setup	108
7.2.4	Document Indexing	108
7.2.5	Short-Text Retrieval Results	109
7.2.6	Long-Text Retrieval Results	111
7.2.7	Discussion	114
7.2.8	Conclusions	115
7.3	Search Prototype	116
7.4	Zero-Shot and Hybrid Retrieval	117

7.4.1	Overview	118
7.4.2	Related Work	120
7.4.3	Methodology	121
7.4.4	Dual Parameters for Hybrid Retrieval	122
7.4.5	Experimental Setup	123
7.4.6	Results and Analysis	124
7.4.7	Discussion	130
7.4.8	Conclusions	132
8	Tetun Users' Search Behavior Analysis	133
8.1	Overview	133
8.2	Related Work	135
8.3	Platforms and Datasets	136
8.3.1	Platforms Overview	136
8.3.2	Data Collection and Datasets	136
8.4	Methodology	137
8.5	Prompt Log Analysis	138
8.5.1	Prompt Length	138
8.5.2	Prompt Frequency	139
8.5.3	Prompt Traffic	140
8.6	User Log Analysis	141
8.6.1	User Evolution	141
8.6.2	User Sessions and Activities	142
8.6.3	User Demographics	143
8.7	Traditional and LLM-Based Search	145
8.7.1	User Search Intent	145
8.7.2	Search Topics and Sequences	146
8.7.3	Prompt/Query Quality Evaluation	148
8.8	LabadainLog-17k+ Dataset Description	148
8.9	Results and Discussion	149
8.10	Limitations and Potential Biases	151
8.11	Conclusions	152
9	Conclusions	154
9.1	Contributions to Tetun IR and NLP	154
9.2	Findings on Tetun Linguistic Aspects	155
9.3	Contributions to Low-Resource Languages IR	155
9.4	Limitations and Open Issues	156
9.5	Future Directions	157
9.6	Final Remarks	157
A	Appendices	158
A.1	Tetun Light Stemmer Algorithm	158
A.2	Qualitative Assessment Criteria	163
A.3	Content Annotation Algorithm	164
A.4	Tetun Stopwords	165
A.5	Search Topic Categories	166
	References	167

List of Figures

1.1	Map of language diversity in Timor-Leste (source: https://vici.mariank-lamer.org/linguistic-maps.html).	2
1.2	Search query logs collected from Timor News.	3
1.3	Phases in the research and development of search solutions for Tetun.	6
2.1	High-level software architecture of an information retrieval system, adapted from Baeza-Yates and Ribeiro-Neto [17].	14
2.2	Learning to rank framework [147].	21
2.3	General neural information retrieval approach for document ranking [161].	22
2.4	Map of Timor-Leste (source: https://www.friendsofossu.org.au/about-ossu-timor-leste/).	27
2.5	The evolution of Tetun.	28
2.6	Examples of basic Tetun phrases.	29
2.7	Examples of Tetun inflection and derivation. The suffix <i>dór</i> is used in both native Tetun and Portuguese loanwords.	30
2.8	Examples of Portuguese loanwords: <i>agradese</i> (<i>agradecer</i> , meaning “thank”), <i>investigasaun</i> (<i>investigação</i> , meaning “research”), <i>krusiál</i> (<i>crucial</i> , meaning “crucial”).	30
4.1	Gaussian-mixture clustering algorithm applied to cluster Tetun (tdt), Portuguese (pt), English (en), and Indonesian (id).	45
4.2	Architecture of the Labadain Crawler.	46
4.3	Document annotation process flow.	53
4.4	Word frequency vs. Word rank.	61
4.5	Evolution of document writing adherence to the Tetun INL standard and the influence of Portuguese loanwords in news articles before and after 2017.	63
5.1	Example of a co-occurrence network generated from the preprocessed text. *English translation: “I am from Timor and I speak Tetun. I live in Porto”.	68
5.2	Top-25 stopwords in Tetun identified using in-degree.	73
5.3	Top-25 stopwords in Emakhuwa identified using in-degree.	74
5.4	Performance of Network-based approaches at mid-range and higher cutoff levels.	78
5.5	Process of constructing a text sample for evaluating the Tetun stemmer’s performance.	81
5.6	UI vs. OI plot showing ERRT distances.	84
6.1	Sample of the document formatted following TREC guidelines: original (left) and English translation (right).	88

6.2	Length distribution of titles and body content in the Labadain-30k+ dataset. In the right-hand figure, the x-axis of the document length distribution is limited to 1,000 tokens (words and numbers) for improved visualization.	89
6.3	Sample of topic formatted following TREC guidelines: original (left) and English translation (right).	90
6.4	Distribution of queries across categories.	91
6.5	General architecture of the retrieval system used for relevance judgments. The document highlighted in <i>red</i> indicates a duplicate that was excluded from the final list after applying the balanced interleaving technique [183].	92
6.6	Web interface for human relevance assessments.	93
6.7	Hourly statistics of the total number of queries evaluated by each assessor.	93
6.8	Total number of relevant and non-relevant documents per query. Relevant documents consist of marginally relevant, relevant, and highly relevant.	95
7.1	Overview of the experimental workflow.	108
7.2	Tetun ad-hoc text retrieval prototype.	116
7.3	Total unique visitors per month in 2024.	117
7.4	Overview of the hybrid retrieval strategy.	119
7.5	Relative performance gains of zero-shot dense retrieval over the baseline.	125
7.6	Relative performance gains of hybrid retrieval over the baseline under the <i>basic</i> configuration ($\alpha = 1.0$, $\beta = 0.3$).	126
7.7	Grid search results for the <i>contextualized</i> configuration combining ColBERTv2 with the Hiemstra LM, evaluated using MAP@10. The combinations $\alpha = 1.1$, $\beta = 0.8$ and $\alpha = 1.5$, $\beta = 1.1$ yield the highest MAP@10 score of 0.2923 (highlighted with red boxes). Since only the top-scoring combination is recorded, the first one encountered during the grid search is selected.	129
8.1	Labadain Chat home page.	134
8.2	Overview of the Methodology.	137
8.3	Top 15 most frequent bigrams in user prompts. *The term “saída” is a misspelling caused by autocorrection, as it is similar to the Portuguese word.	140
8.4	Average distribution of total and unique prompts throughout the day.	141
8.5	Total of unique new and returning users per day.	142
8.6	Total number of users and unique users per hour.	143
8.7	Time spent by users per session.	144
8.8	User intent from prompt analysis by human annotators across three platforms: Labadain Chat (conversational search), Labadain Search (keyword search), and Timor News (keyword search from Google). The “yes” indicates the presence of the respective class in the prompt (i.e., interrogative, imperative, modal, or pronoun), while “no” indicates its absence.	145
8.9	Visualization of association rules in user follow-up search sequences. Edges represent relationships between themes, with arrow directions indicating the sequence of actions. Edge labels denote lift values, and the red arrow highlights the strongest association with a lift of 1.31.	148

List of Tables

4.1	Summary of the dataset used for training and evaluating the LID model.	44
4.2	Accuracy of the models' performance when evaluating using the development set.	45
4.3	Top five domains in the dataset.	49
4.4	Summary of the corpus content (excluding titles and URLs). *Tokens do not include punctuation and special characters.	49
4.5	Total number of text pages per data source.	50
4.6	Total number of text pages per top-level domain.	50
4.7	Results of the corpus quality assessment.	51
4.8	Summary of the web-crawled dataset.	55
4.9	Summary of the final dataset.	57
4.10	Comparison of the Labadain-30k+ dataset size and total number of speakers with other low-resource languages.	58
4.11	Description of the Labadain-30k+ Dataset.	59
4.12	Summary of the documents in the Labadain-30k+ dataset.	60
4.13	Top five sources by document count.	60
4.14	Summary of the dataset per TLDs.	60
4.15	Yearly evolution of document quantity. +Percentage points.	62
4.16	Evolution of Tetun INL adoption in document writing before and after 2017. +Numbers are excluded from the count. *Some loanwords are also included in the Tetun INL dictionary.	62
5.1	Summary of the Dataset. Each document corresponds to a text line separated by a newline character.	67
5.2	Feature ranking with information gain.	71
5.3	Performance of stopword detection methods for English.	72
5.4	Performance of stopword detection methods for Portuguese.	72
5.5	Performance of stopword detection methods for Tetun and Emakhuwa	73
5.6	Summary of Tetun, Portuguese, and English datasets used in stopword detection.	76
5.7	Stopword precision for Tetun.	76
5.8	Stopword precision for Portuguese.	76
5.9	Stopword precision for English.	77
5.10	Transformation rules for Portuguese-derived words applied to Tetun suffixes.	79
5.11	Summary of words produced at each stage of the sample word generation process. *Approximately 88.4% reduction in the number of words was observed after applying the LID model, attributed to the model's average score per word of 0.394, with a threshold set to 0.95.	81
5.12	Cohen's k score for inter-annotator agreement in constructing the ground truth set.	82
5.13	Summary of the ground truth set and word length.	82

5.14	Analysis of the stemming algorithms' performance using the Paice metric.	84
6.1	Examples of original query logs and their reformulations. Words highlighted with a <i>green</i> background indicate newly added terms, while word with an <i>orange</i> background indicates orthography correction.	90
6.2	Summary of queries.	91
6.3	Cohen's kappa scores for inter-annotator agreement among the five assessors. . .	94
6.4	Summary of documents with tied scores.	94
6.5	Details of the human judgment results.	95
6.6	Summary of the final test collection. * Relevant documents consist of marginally relevant, relevant, and highly relevant.	95
6.7	Comparison of the <i>Labadain-Avaliadór</i> with other LRLs.	96
6.8	Price associated with LLMs as of May 2024.	99
6.9	Examples of Tetun to English translations provided by LLMs.	99
6.10	Examples of query-document pairs.	99
6.11	Summary of document lengths (in words).	100
6.12	Evaluation of translation quality using BLEU.	100
6.13	LLM capabilities of Multi-task Language Understanding on MMLU.	100
6.14	Cohen's Kappa scores measuring the agreement between human assessors and LLMs.	102
6.15	Comparison of inter-annotator agreement levels in relevance judgment tasks. . . .	102
6.16	Expense on LLMs for relevance judgment tasks.	103
7.1	The evaluation results.	106
7.2	Index compression factor (%) for titles and content compared to the baseline. . .	109
7.3	Effectiveness of text preprocessing techniques in short-text retrieval. Red values indicate scores lower than the baseline.	110
7.4	Effectiveness of combined text preprocessing techniques in short-text retrieval. Values highlighted with a <i>green</i> background indicate the best score at the respective metric and cutoff.	111
7.5	Effectiveness of text preprocessing techniques in long-text retrieval. Red values indicate scores lower than the baseline.	112
7.6	Effectiveness of combined text preprocessing techniques in long-text retrieval. Values highlighted with a <i>green</i> background indicate the best score at the respective cutoff.	113
7.7	Performance of contextualized hybrid retrieval in Tetun. Scores in bold indicate improvement over the baseline. Values with a <i>green</i> background represent the best results ($p < 0.05$). The symbol [†] indicates that the model outperformed its counterpart (monolingual vs. multilingual).	128
7.8	Statistics for queries and documents containing Portuguese loanwords. Loanwords were identified by matching each preprocessed word against the Portuguese loanword dictionary published by Greksáková [103]. * Tokens include words and numbers.	131
8.1	Sample of interaction logs from Labadain Chat.	136
8.2	Prompt length characteristics.	139
8.3	Summary of user prompts and sessions.	142
8.4	Comparison of search intent across conventional and LLM-based platforms. . . .	146
8.5	Comparison of search topics between conventional and LLM-based platforms. . .	147

8.6	Examples of log data entries in LabadainLog-17k+.	149
8.7	Key similarities and differences between this study and Trippas et al. [218].	150
A.1	Linguistic regions used in the <i>Labadain-Stemmer</i>	158
A.2	List of Portuguese-derived suffixes.	159
A.3	Assessment criteria used for the qualitative evaluation of the text corpus.	163
A.4	Tetun stopwords list (<i>Labadain-Stopwords</i>) with English translations.	165
A.5	Descriptions of search topic categories.	166

Abbreviations and Symbols

BERT	Bidirectional Encoder Representations from Transformers
BM25	Okapi Best Matching
CC	Common Crawl
CLEF	Conference and Labs of the Evaluation Forum
DFR	Divergence From Randomness
DIT	Dili Institute of Technology
GPT	Generative Pre-trained Transformer
IDF	Inverse Document Frequency
INL	Instituto Nacional de Linguística (National Institute of Linguistics)
IR	Information Retrieval
LID	Language Identification
LM	Language Model
LLM	Large Language Model
LRL	Low-Resource Language
MAP	Mean Average Precision
NDCG	Normalized Discounted Cumulative Gain
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
TF	Term Frequency
TF-IDF	Term Frequency-Inverse Document Frequency
TREC	Text REtrieval Conference
WWW	World Wide Web

Chapter 1

Introduction

Information retrieval (IR) systems are essential tools for searching and accessing information on the internet and digital platforms. Search systems like Google and Bing have empowered users to find and consume information efficiently, supporting tasks such as locating academic resources, staying updated with news, or accessing government services. These systems have revolutionized information access, enabling users to find the information they need effectively.

Despite these advancements, the digital shift has introduced significant challenges for low-resource languages (LRLs), as the dominant web content is in English. This imbalance creates barriers to accessing the content in LRLs, often resulting in irrelevant search results, user frustration, and unsatisfactory search experiences [32, 36, 114]. The scarcity of digital resources for underrepresented languages reinforces digital inequality and limits the participation of their communities in the broader online ecosystem.

Tetun, one of Timor-Leste’s official languages [226], faces similar challenges. Timor-Leste, an island country in Southeast Asia, has a population of 1.18 million, with approximately 79% reported as Tetun speakers [57, 62]. Despite being spoken by most of the population, Tetun remains underrepresented on the web, with limited digital content available as of 2023 [62, 136]. This scarcity, coupled with the lack of search algorithms and techniques tailored to Tetun, often leads to irrelevant search results for text-based natural language queries. These limitations hinder the Tetun-speaking community from accessing the information they need, highlighting a critical gap in digital inclusion and equitable access to search technologies.

To address these challenges, it is crucial to develop IR algorithms and techniques tailored specifically to handle the linguistic characteristics of Tetun. Such innovations would improve access to information in Tetun, help preserve its linguistic heritage, and elevate its status in the digital landscape. Advancements in IR-based solutions would also benefit key sectors such as governance, health, education, and the judicial system, enabling the Timorese community to overcome language barriers and access information in their language. By fostering digital inclusivity, these efforts would significantly contribute to the country’s overall development.

1.1 Context and Motivation

Timor-Leste is a multilingual country shaped by its complex history, including Portuguese colonial rule from 1702 to October 1975 and the struggle for independence following Indonesia’s invasion from December 1975 to September 1999 [104]. After restoring its independence in May 2002, the country designated Tetun and Portuguese as its official languages and Indonesian and English as working languages [226]. Additionally, over 30 dialects are spoken across the country, reflecting its rich linguistic diversity [57, 60]. The municipal distribution of most of these languages is illustrated in Figure 1.1.

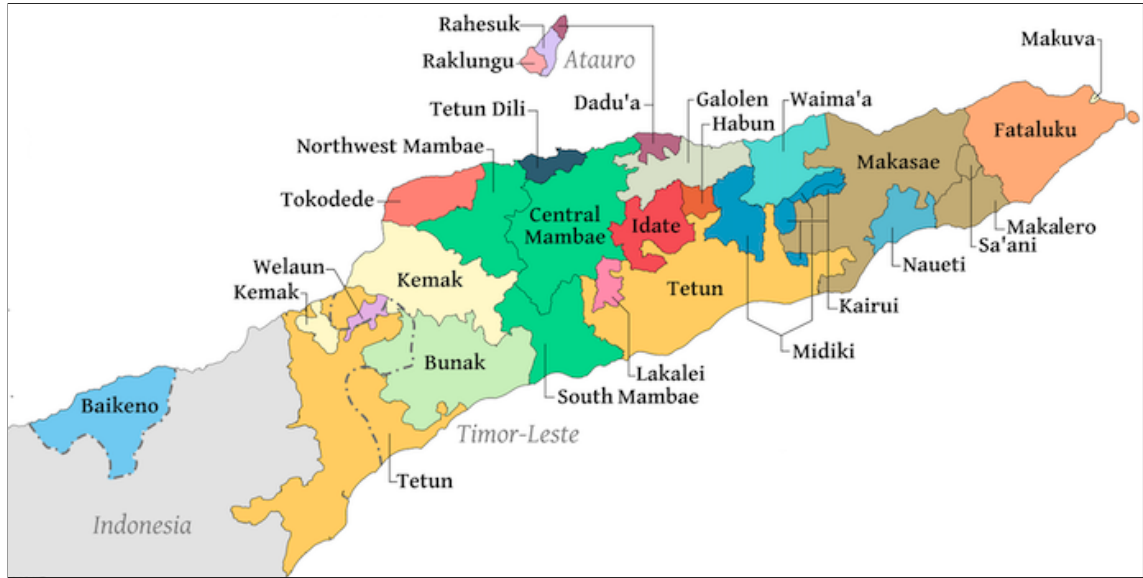


Figure 1.1: Map of language diversity in Timor-Leste (source: <https://vici.marianklamer.org/linguistic-maps.html>).

There are two main varieties of Tetun: Tetun Dili or Tetun *Prasa* (commonly referred to as Tetun) and Tetun Terik [224]. Tetun Dili comprises two standardized forms, Tetun *Instituto Nacional de Linguística* (INL) and Tetun Dili Institute of Technology (DIT), while Tetun Terik remains one of Timor-Leste’s dialects. Tetun INL is the official Tetun standardized by the government through the INL, whereas Tetun DIT is developed by linguists at DIT and differs from Tetun INL in certain orthographic conventions [224]. Tetun INL is the primary language used in the education system, official publications, and media [169].

Logs of real-world searches conducted by Tetun-speaking users and captured by the Timor News platform¹ reveal recurring issues: search queries are often improperly constructed and frequently mix Tetun with other languages commonly spoken in Timor-Leste, such as Portuguese and Indonesian, as shown in Figure 1.2. This linguistic mix leads to poorly structured queries, making it more challenging for existing search systems to retrieve relevant results for the given query.

IR systems are essential tools for addressing such challenges, as they enable access to information across diverse digital platforms and the broader web through search systems. One of

¹<https://www.timornews.tl>

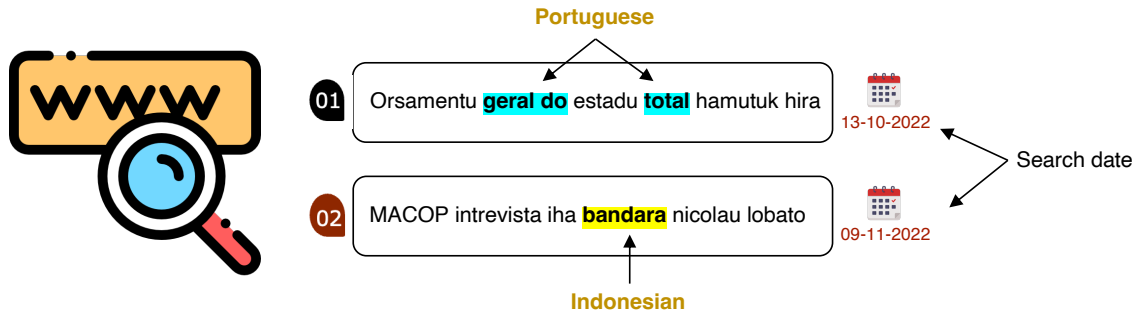


Figure 1.2: Search query logs collected from Timor News.

the most notable IR applications is keyword-based web search, as showcased by Google, where users submit natural language text queries to find relevant information. However, existing IR systems are primarily designed for high-resource languages, such as English, leaving languages like Tetun underserved. This lack of support intensifies the difficulties faced by Tetun-speaking users, who often struggle to find relevant digital content due to the absence of search solutions designed particularly to handle their linguistic characteristics.

Developing search solutions tailored specifically to Tetun is essential for bridging this digital divide. Such solutions can empower the Tetun-speaking community to fully engage with the digital revolution by enabling equitable access to information in their native language. These advancements would not only support educational and professional development but also strengthen cultural identity and inclusivity in the digital era.

1.2 Problem Description

Tetun, historically used as a lingua franca in trade and by the church, became one of Timor-Leste's official languages when the country restored its independence in May 2002. Despite its official status, digital content in Tetun remains scarce. Notably, growth in digital content has only been observed in 2017, largely driven by online news portals [62, 136]. This limited availability of digital resources has constrained the development of language technology for Tetun.

In the era of the digital revolution, accessing information in a language that people can understand is increasingly crucial, including for Tetun speakers. Addressing this need requires effective retrieval solutions that can handle the language-specific nuances of Tetun. However, several challenges hinder the development of such solutions, including data scarcity, language standardization issues, and limited funding and research initiatives.

Another major challenge arises from Timor-Leste's multilingual environment, which contributes to the prevalence of non-standardized Tetun, particularly in its written form. This linguistic diversity often leads to language mixing in both spoken and written communication. Additionally, Tetun's evolution is shaped by its history of colonization and language development. The earliest written record of Tetun was documented by a Portuguese priest [103, 223], with Portuguese

orthography directly applied to the language. Consequently, Tetun incorporates a substantial number of Portuguese loanwords, particularly in verbs, nouns, and adjectives [103, 106, 223]. This Portuguese influence is especially prominent in media such as newspapers and television in the post-independence era. Furthermore, Indonesian has also left its mark, leading to a highly variable written form of Tetun influenced by both Portuguese and Indonesian.

Given that Tetun was a dialect and later became one of the official languages, the lack of text corpora and information retrieval techniques tailored to its linguistic characteristics remains a significant barrier. This absence limits the ability to develop effective search solutions for Tetun text-based queries.

Problem Statement: *The core challenge in Tetun text information retrieval lies in the unavailability of Tetun text corpora and the lack of IR techniques specifically designed to address Tetun's unique linguistic characteristics, which hinders the effective retrieval of documents that meet information needs in Tetun text-based searches.*

The aforementioned problems highlight the need for innovative solutions to bridge the gap between Tetun's linguistic complexities and the demands of information retrieval systems. Tackling these issues requires developing foundational language resources to support such efforts. These initiatives are essential for enhancing access to information in Tetun, strengthening its digital presence, and empowering its speakers within the online ecosystem.

1.3 Objectives

This thesis aims to develop search solutions for Tetun, with a specific focus on investigating the ad-hoc text retrieval task. The work addresses the language-specific challenges associated with Tetun to create effective retrieval solutions tailored to Tetun text-based search.

To achieve this goal, the research concentrates on the following key activities:

- **Developing preprocessing tools and algorithms** tailored to Tetun to handle its unique linguistic characteristics.
- **Constructing text resources for Tetun**, which are essential for supporting text retrieval tasks and foundational linguistic research.
- **Investigating retrieval strategies** that account for Tetun's linguistic characteristics, enabling the development of effective search solutions.
- **Evaluating the impact of linguistic features on retrieval effectiveness** to identify strategies that enhance performance in Tetun text-based search.

By achieving these objectives, this research enables access to digital information in Tetun and contributes to the advancement of IR for LRLs. Additionally, it provides methodologies for search solutions that can be adapted by other underrepresented languages posing similar challenges.

1.4 Research Questions and Hypothesis

To guide the development of search solutions for Tetun, this thesis investigates the following key research questions (RQs):

RQ1: *How can a text corpus and a test collection be built for Tetun, considering its low-resource status?*

Developing a text corpus for Tetun requires addressing challenges associated with data scarcity in a multilingual context. This research question aims to explore approaches for constructing text corpora for LRLs that can be adapted to build a Tetun text corpus and applied to other LRLs facing similar challenges. The resulting dataset serves as a foundational resource for Tetun, specifically enabling the development of a test collection for the ad-hoc text retrieval task.

RQ2: *How can text preprocessing techniques tailored to Tetun’s linguistic characteristics improve retrieval effectiveness?*

Tetun words include accented letters (á, é, í, ó, ú, ñ), apostrophes, and hyphens, and their writing form is shaped by the multilingual environment in Timor-Leste. This research question aims to investigate the extent to which query and document preprocessing techniques affect the retrieval effectiveness of Tetun text-based searches.

RQ3: *What strategies provide the most effective solutions for Tetun text-based search?*

Text retrieval strategies for ad-hoc tasks encompass a variety of approaches, with studies reporting inconsistent effectiveness across languages, strategies, and models. These approaches often involve preprocessing techniques such as stopword removal and stemming. This research question aims to investigate which retrieval strategies are most effective for Tetun text-based search.

To respond to these research questions, we propose the following hypothesis:

Hypothesis: *Applying text preprocessing techniques tailored to the linguistic characteristics of Tetun can improve the effectiveness of ad-hoc retrieval, particularly for short texts without stemming.*

This hypothesis is grounded in initial findings on ad-hoc text retrieval for Tetun, which showed that indexing document titles without stemming delivered the best performance, with a 3.1% relative gain in overall MAP and 1.9% in overall NDCG compared to stemming [61].

1.5 Methodology and Key Contributions

To address the research questions and validate the hypothesis, we began with the development of the research idea, which was validated through two publications, formalized in a thesis proposal, and subsequently defended. Following this, a data collection pipeline was designed and employed to acquire Tetun text documents from the World Wide Web (WWW). These documents were annotated by native Tetun speakers to create a Tetun text dataset. This dataset served as the foundation for building a stopwords list, designing a stemmer, and developing a test collection for evaluation.

We then conducted experiments with various retrieval strategies to establish baselines for Tetun ad-hoc text retrieval and showcased the most effective strategy in a search prototype. Building on this foundation, we further explored neural-based retrieval approaches and the application of large language models (LLMs) to Tetun ad-hoc text retrieval. Finally, to gain a deeper understanding of the Tetun-speaking users' search behavior, we analyzed real-world search log data collected from two traditional search platforms and a conversational search platform developed specifically for the Tetun-speaking community. **Figure 1.3** presents the structured progression of research and development for Tetun search solutions in different phases. The focus and contribution of each phase are described in detail below.

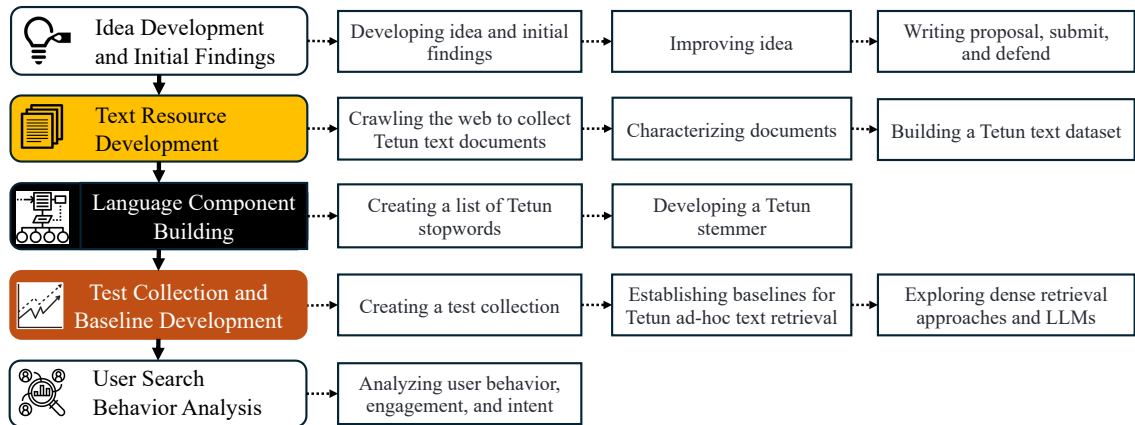


Figure 1.3: Phases in the research and development of search solutions for Tetun.

Idea Development and Initial Findings This initial stage involved formulating research ideas, refining concepts, preparing the research proposal for submission and defense, and formalizing its objectives to establish the foundation of the work. Two papers were published during this stage: the first at the 10th Ph.D. Symposium on FDIA in 2022 [61], which presented the initial research idea and preliminary findings on Tetun ad-hoc text retrieval, and the second at the Doctoral Consortium at ECIR 2023 [60], which provided an improved version of the research idea. Feedback from these publications informed the finalization, submission, and defense of the research plan.

Text Resource Development With a clear research direction, this stage focused on collecting and constructing a Tetun text dataset. Key activities included web crawling to gather Tetun text

documents and constructing a dataset, and characterizing the collected documents to ensure representation across broad topics and diverse sources. This phase resulted in the publication of two papers [62, 74], with key contributions including a crawler pipeline designed specifically for LRLs, a Tetun tokenizer, a Tetun language identification model, and a Tetun text dataset suitable for a variety of IR and natural language processing tasks.

Language Component Building In this stage, the focus shifted to developing essential preprocessing algorithms and language components tailored to Tetun. These included a stopword list and a stemmer. These components form the foundation for investigating effective text processing and retrieval in Tetun. The key outputs were a novel approach for stopwords detection, a Tetun stopword list, and a Tetun stemmer with three variants. The outcomes of this work were reported in two research papers: one published [10], and another submitted to a journal and under review [73].

Test Collection and Baseline Development This stage aimed to enable the evaluation of IR techniques for Tetun and to establish baselines for ad-hoc text retrieval tasks. Key activities included creating a test collection and experimenting with various retrieval strategies to identify the most effective approach for Tetun text retrieval. The test collection provides a foundation for assessing retrieval effectiveness, while the baselines serve as reference points for measuring future improvements. Additionally, advanced retrieval techniques, such as zero-shot dense and hybrid retrieval strategies leveraging LLMs, were explored and compared with the sparse baseline, establishing an additional dense retrieval baseline for Tetun ad-hoc text retrieval. The outcomes of this phase, including the test collection and baselines for Tetun ad-hoc retrieval, were reported in three research papers: two published [65, 77] and one under review [73].

User Search Behavior Analysis This final phase examined the search behavior, engagement, and intent of the Tetun-speaking community. Key activities involved analyzing user queries to identify common patterns, focusing on their characteristics, intents, and topics. The findings were reported in a research paper [125].

1.6 Publications and Research Outputs

The following sections present the published and under-review papers, together with the datasets, tools, and algorithms developed as part of this thesis.

1.6.1 Published Papers

The original papers published as part of this thesis are listed below:

- [1] **Gabriel de Jesus** and Sérgio Nunes (2025). Insights into LLM-Based Conversational Search: A Study of Tetun-Speaking Users' Search Behavior. *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information*

- Retrieval (ICTIR)*, 297–306, Padua, Italy, 2025. URL <https://doi.org/10.1145/3731120.3744596>.
- [2] **Gabriel de Jesus**, Siddharth A.K. Singh, Sérgio Nunes, and Andrew Yates (2025). Zero-Shot and Hybrid Strategies for Tetun Ad-Hoc Text Retrieval. *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, 264–274, Padua, Italy, 2025. URL <https://doi.org/10.1145/3731120.3744593>.
- [3] **Gabriel de Jesus** and Sérgio Nunes (2024). Exploring Large Language Models for Relevance Judgments in Tetun. In Siro et al. (eds.), *Proceedings of The First Workshop on Large Language Models for Evaluation in Information Retrieval (LLM4Eval 2024)*, co-located with 47th International ACM SIGIR Conference, vol 3752, 19–30, Washington D.C., USA, July 18, 2024. URL <https://ceur-ws.org/Vol-3752/paper2.pdf>.
- [4] **Gabriel de Jesus** and Sérgio Nunes (2024). Data Collection Pipeline for Low-Resource Languages: A Case Study on Constructing a Tetun Text Corpus. In Calzolari et al. (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20–25 May 2024, Torino, Italy*, pages 4368–4380. ELRA and ICCL, 2024. URL <https://aclanthology.org/2024.lrec-main.390>.
- [5] **Gabriel de Jesus** and Sérgio Nunes (2024). Labadain-30k+: A Monolingual Tetun Document-Level Audited Dataset. In Melero et al. (Eds.), *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 177–188, Torino, Italia, May 2024. ELRA and ICCL. URL <https://aclanthology.org/2024.sigul-1.22>.
- [6] Felermينو D. M. A. Ali, **Gabriel de Jesus**, Henrique Lopes Cardoso, Sérgio Nunes, and Rui Sousa-Silva (2024). Network-Based Approach for Stopword Detection. In Gamallo et al. (Eds.), *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 2*, 55–63. Association for Computational Linguistics, Santiago de Compostela, Spain. URL <https://aclanthology.org/2024.propor-2.9>.
- [7] **Gabriel de Jesus** (2023). Text Information Retrieval in Tetun. In: Kamps, J., et al. (Eds.), *Proceedings of the 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023*, Advances in Information Retrieval, Part III. Lecture Notes in Computer Science, vol. 13982. Springer. URL https://doi.org/10.1007/978-3-031-28241-6_48.
- [8] **Gabriel de Jesus** (2022). Text Information Retrieval in Tetun: A Preliminary Study. *The 10th edition of the PhD Symposium on Future Directions in Information Access (FDIA)*, July 20, 2022, Lisbon, Portugal. URL <https://doi.org/10.48550/arXiv.2406.07331>.

1.6.2 Paper Under Review

The following paper has been submitted to the ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) and is under review:

- [1] **Gabriel de Jesus** and Sérgio Nunes (2025). Establishing a Foundation for Tetun Ad-Hoc Text Retrieval: Stemming, Indexing, Retrieval, and Ranking. *ACM Transaction Asian Low-Resource Language Information Processing (TALLIP)*. Association for Computing Machinery, New York, NY, USA, 2025. Pre-print URL <https://doi.org/10.48550/arXiv.2412.11758>

1.6.3 Datasets, Tools, Algorithms, Software, and Prototypes

Tetun datasets, algorithms, tools, and prototypes developed as part of this thesis are listed below.

Datasets

- [1] *Labadain-30k+*: A monolingual Tetun text dataset comprising 33,550 documents from June 2001 to September 2023 [64].
- [2] *Labadain-Avaliadór*: A Tetun test collection for ad-hoc text retrieval consisting of 59 topics, 33,550 documents, and 5,900 query-document relevance judgments (*qrels*) [70].
- [3] *Labadain-Stopwords*: A curated list of 160 Tetun stopwords, compiled from the Labadain-30k+ dataset and validated by native Tetun speakers [72].
- [4] *LabadainLog-17k+*: A log dataset compiled from three platforms: Labadain Chat, Labadain Search, and Google Search Console for Timor News. It contains 16,952 conversational prompts from Labadain Chat, 400 queries from Labadain Search, and 400 queries from Google Search Console [71].
- [5] *Labadain-ZSRunS*: A dataset consisting of run files produced by classical sparse and zero-shot dense retrieval models applied to Tetun ad-hoc text retrieval, together with document summaries generated by an LLM from the content of each document in the *Labadain-Avaliadór* collection [76].

Algorithms, Tools, Software, and Prototypes

- [1] *Labadain-Stemmer*: A stemming algorithm with three variants for Tetun [75].
- [2] *Tetun Tokenizer*: A Python package for segmenting Tetun text into tokens [67].
- [3] *Tetun LID*: A machine learning model for language identification in text documents, supporting Tetun, Portuguese, English, and Indonesian [66].

- [4] *Labadain Crawler*: A data collection pipeline for crawling textual documents from the web for low-resource languages [63].
- [5] *Labadain Search*: A web-based search engine prototype for showcasing ad-hoc text retrieval in Tetun [69].
- [6] *Labadain Chat*: A web-based, LLM-powered conversational search prototype for collecting Tetun-speaking users' search log data [68].

1.7 Thesis Structure

The remaining chapters of the thesis are organized as follows.

Chapter 2: Background This chapter introduces the key concepts of information retrieval (IR) relevant to the thesis, including an overview of IR and its components, methodologies for constructing test collections, retrieval and ranking models, and evaluation metrics. It also provides an overview of Tetun and its linguistic evolution, focusing on orthography, morphology, loanwords, and the availability of Tetun text documents on the web as of 2023.

Chapter 3: Text information retrieval in Low-Resource Languages This chapter reviews the existing literature on text IR in low-resource languages (LRLs), discussing the key characteristics of LRLs, approaches to constructing datasets and test collections, and retrieval and ranking experiments relevant to the objectives of this thesis.

Chapter 4: Tetun Text Resource Development This chapter introduces the Labadain Crawler, the Labadain-30k+ dataset, and an analysis of its composition and evolution on the web. The crawler was used to collect Tetun documents from the web, which were then compiled into the Labadain-30k+ dataset.

Chapter 5: Tetun Language Component Building This chapter presents a novel approach to stopword detection based on the topological properties of co-occurrence networks, applied to generate a stopword list for Tetun. It also describes the development of a Tetun stemmer with three variants and reports the results of its intrinsic evaluation.

Chapter 6: Tetun Test Collection This chapter describes the step-by-step process of constructing a Tetun test collection and presents findings from experiments with large language models (LLMs) for automated relevance judgments.

Chapter 7: Tetun Ad-Hoc Text Retrieval This chapter presents experiments on ad-hoc text retrieval in Tetun. It begins with findings from a preliminary study, followed by results that establish foundational baselines. The chapter then reports results from experiments with pretrained dense retrieval models using zero-shot and hybrid strategies.

Chapter 8: Tetun Users' Search Behavior Analysis This chapter examines the search behavior, engagement, and intent of Tetun users based on interaction logs with an LLM-based conversational search system. It then compares these findings with those of English-speaking users to identify similarities and differences between the two communities. Finally, it investigates the influence of the LLM-based conversational system on user behavior in conventional search engines.

Chapter 9: Conclusions This chapter summarizes the key contributions and findings of this thesis and highlights its limitations, open issues, and potential directions for future work.

Chapter 2

Background

This chapter provides fundamental background on the core concepts of information retrieval (IR) and Tetun relevant to this thesis. It begins with an overview of foundational IR concepts, including basic principles and components, test collections, text retrieval and ranking models, and evaluation metrics. The focus then shifts to Tetun, outlining its linguistic evolution and role in newspapers. Furthermore, the chapter discusses key linguistic features, including orthography, morphology, Portuguese loanwords, and the availability of Tetun text documents on the web as of 2023.

2.1 Information Retrieval

The first motivation for developing information access technologies arose from the rapid growth of scientific publications after World War II [145]. The idea of using computers to automate the indexing and searching of relevant information gained popularity with Vannevar Bush’s article “*As We May Think*”, published by the Atlantic in 1945 [34].

According to Sanderson and Croft [195], the computer capable of searching for information was described by Holmstrom in 1948, and the term “information retrieval” was first used by Calvin Mooers in 1950. The systematic research of IR began in 1957 when Cyril Cleverdon initiated the Cranfield project, which conducted a series of experimental studies of IR on large document collections and established test collections and evaluation metrics for IR systems [194].

In 1992, the National Institute of Standards and Technology (NIST) and the U.S. Information Technology Office of the Defense Advanced Research Projects Agency (DARPA) co-sponsored the Text REtrieval Conference (TREC)¹ as part of the TIPSTER Text program. This initiative was established to advance IR research by introducing a standardized evaluation framework, enabling large-scale experimentation, and fostering collaboration within the research community [17]. TREC served as a model for subsequent large-scale evaluation campaigns, including the NII Testbeds and Community for Information Access Research (NTCIR), the Conference and Labs of the Evaluation Forum (CLEF), and the Forum for Information Retrieval Evaluation (FIRE).

¹<https://trec.nist.gov/overview.html>

NTCIR² was launched in 1998 by Japan’s NII to focus on East Asian languages, promoting tasks such as cross-lingual IR, web search, and question answering. CLEF,³ founded in 2000, aims to advance multilingual and multimodal information access in European languages through its labs and workshops, contributing to the development of domain-specific retrieval solutions. FIRE,⁴ launched in 2008 in India, focuses on advancing IR research for low-resource South Asian languages, such as Hindi, Bengali, and Tamil, and has expanded to include cross-lingual IR and machine translation. These evaluation initiatives have been foundational in advancing IR research by establishing standardized datasets, shared tasks, and reproducible evaluation methodologies, empowering the research community to benchmark progress and validate system performance.

2.1.1 Definition

Gerard Salton introduced the first definition of information retrieval as a field concerned with the structure, analysis, organization, storage, search, and retrieval of information [257]. While multiple definitions of IR exist, the one by Manning et al. [158] is particularly relevant to this thesis, given its focus on unstructured text retrieval:

“Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)”.

2.1.2 Ad-Hoc Text Retrieval

Ad-hoc text retrieval is the task of retrieving documents from large text collections in response to a specific user query, without prior knowledge of the topics users are likely to search, representing the unpredictable nature and short duration of each search session [229, 232]. The ad-hoc retrieval task was a central focus of the TREC tracks held from 1992 to 1999 and was revisited on the robust track from 2003 to 2005 [158, 231].

The concept of the ad-hoc retrieval task is to assess a system’s ability to retrieve and rank relevant documents for previously unseen queries, simulating real-world search scenarios. It emphasizes the system’s capacity to respond to new information needs without relying on predefined query-document relationships.

The ad-hoc track laid the foundation for standards that influenced many subsequent TREC tracks and larger IR evaluation efforts [17, 194]. The three essential components of a test collection to evaluate the effectiveness of an IR system are the document collection, *topics* (queries), and *qrels* (query-document relevance judgments).

²<http://research.nii.ac.jp/ntcir/index-en.html>

³<https://www.clef-initiative.eu>

⁴<http://research.nii.ac.jp/ntcir/index-en.html>

2.1.3 Architecture

The classical high-level software architecture of an IR system comprises four main components: crawling, indexing, retrieval, and ranking [17], as shown in Figure 2.1. The process begins with a document collection, typically obtained through web crawling, where the collected documents are stored in a central repository and indexed using an inverted index. The inverted index organizes distinct terms in the collection and maps each term to the documents in which it appears. Once the indexing is completed, the retrieval process can be initiated.

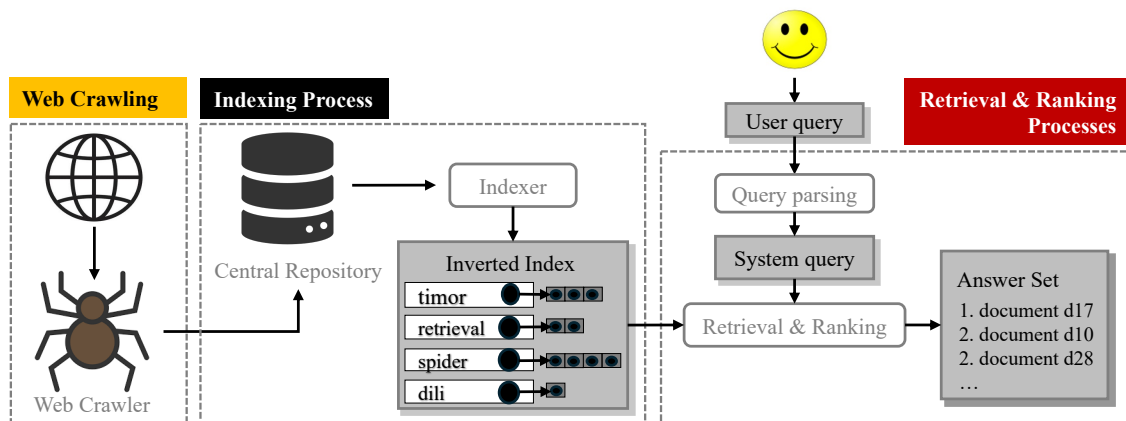


Figure 2.1: High-level software architecture of an information retrieval system, adapted from Baeza-Yates and Ribeiro-Neto [17].

In the retrieval and ranking phase, users search by submitting a natural language text query that represents their information needs. The query undergoes preprocessing, such as stopwords removal and stemming, to enhance its representation before matching with the document index to retrieve a subset of relevant documents. The same preprocessing steps are applied to both documents and queries, with preprocessing for documents performed before indexing. Finally, the retrieved documents are ranked, and the top- k results are presented to users in a list ordered by relevance in response to their search, where the most relevant documents are at the top of the list, with less relevant ones further down.

2.1.4 Web Crawling

Web crawling is a widely used technique for constructing document collections, particularly by search engines such as Google, Bing, and Yahoo, to gather content from the World Wide Web (WWW) [7]. This method employs a web crawler to systematically browse the web and store the retrieved documents in a central repository for indexing [113, 163].

A web crawler, also known as a spider or robot, is a program, software, or script designed to automate this task systematically and efficiently. The general structure of a web crawler consists of three main components [163]: a *frontier* that maintains a list of URLs to visit, a *page downloader* responsible for obtaining pages from the WWW, and a *web repository*, which stores the retrieved web content in a database.

Several free and open-source web crawler tools are available for collecting web documents, such as Apache Nutch, Heritrix, and StormCrawler. Apache Nutch⁵ is an open-source web crawler written in Java that offers a scalable and extensible framework for crawling, retrieving, and processing web content. It is built on top of Apache Hadoop and leverages its distributed data structures to support large-scale web crawling. Nutch integrates various components for indexing, metadata extraction, and advanced IR tasks.

Similarly, Heritrix⁶ is an open-source web crawler, also implemented in Java, developed in 2003 by the Internet Archive in collaboration with the Nordic national libraries. Unlike Nutch, Heritrix is specifically designed for web archiving and preservation, with features tailored to capture and store web content at scale. StormCrawler,⁷ on the other hand, is conceptually similar to Nutch but is built on top of Apache Storm. It supports real-time, distributed web crawling and processing, making it well-suited for applications that require continuous data ingestion and low-latency processing from the web.

2.1.5 Text Preprocessing

Text preprocessing is a technique that transforms the original text into a standardized format to reduce query-document mismatches and improve retrieval effectiveness. To ensure accurate matching, both the document and the query text must undergo the same preprocessing steps [55]. Common text preprocessing techniques include language identification, text normalization, tokenization, stopword removal, and stemming.

Language Identification Language mixing is a common challenge in web document collections and multilingual environments. Language identification (LID) techniques address this challenge by automatically detecting specific languages within a given text, including documents that contain multiple languages. To advance research in this area, CLEF introduced the LogCLEF track [157] in 2009, which focused on classifying user search queries by language and adapting retrieval strategies to better support multilingual search contexts.

Text Normalization Text normalization involves transforming text into a standardized representation to enhance the accuracy, efficiency, and effectiveness of downstream tasks, such as text retrieval. This process typically includes converting text to lowercase and removing special characters and punctuation. Lowercasing, also called case-folding, is the text transformation technique that addresses sparsity issues by standardizing word capitalization, which is generally not important in search and has minimal impact on retrieval effectiveness [55]. Lowercase converts words with different cases into the same lowercase form. For example, the words [Timor, TiMOR, TiMoR, TiMOR] are all transformed to “timor”.

⁵<https://nutch.apache.org>

⁶<https://github.com/internetarchive/heritrix3>

⁷<http://stormcrawler.net/index.html>

Tokenization Tokenization is a fundamental preprocessing step in NLP and IR, where texts are segmented into smaller units, known as *tokens*, to facilitate text processing tasks. Traditionally, in Latin-based script languages, tokens are broadly defined as contiguous sequences of non-punctuation, non-whitespace characters or as individual punctuation marks [160]. For specific punctuation marks, such as hyphens and apostrophes, tokenization decisions are guided by language-specific characteristics to produce tokens that approximate word forms effectively. In the context of IR, Croft et al. [55] highlighted the impact of the tokenization process in search, emphasizing that the same tokenization techniques must be applied to both queries and documents to ensure alignment between the index terms in the documents and those in the queries.

Stopword Removal Stopwords are function words (such as articles, prepositions, and conjunctions) that appear frequently in documents but generally provide minimal value to retrieving relevant documents for a given query [17, 158]. Removing these words from both queries and documents can improve retrieval efficiency and effectiveness [17, 55]. Traditionally, stopwords lists are created by selecting the most frequent terms in a corpus, typically choosing the top-*n* most common terms (e.g., the top 50) [55, 158]. Manning et al. [158] observed shifts in IR systems over time, moving from the use of relatively large stopwords lists (200-300 terms) to very small ones (7-12 terms) and, in some cases, to not using stopwords at all. In phrase searches, however, stopwords play a crucial role, as all search terms must appear in the documents in the same adjacent order as the user's query.

Stemming Stemming is an essential component of text processing that captures the relationship between different variations and forms of a word resulting from inflection (e.g., plurals, tenses, and gender) or derivation (e.g., converting a verb into a noun by adding suffixes) and reduces them to a common root [55]. A stem is the root form of a word that remains after the removal of its affixes [17]. For example, the words [car, cars, car's, cars'] are all reduced to "car". The first stemming algorithm was proposed by Lovins [150], based on the principles of iteration and longest match. The iteration principle assumes that affixes are attached to stems in a particular order from a predefined set of affixes. The algorithm removes affixes from either the beginning or the end of the word, depending on which affix class is detected. According to the longest match principle, if multiple endings within a class match, the longest one should be removed. Since then, various stemming techniques have been developed, including rule-based, dictionary-based, and machine learning-based approaches. A notable example is the Porter Stemmer [178], a widely used rule-based suffix removal algorithm for English, recognized for its simplicity and performance [17].

2.1.6 Indexing, Retrieval, and Ranking

Various indexing techniques are used in IR, with the inverted index being the most widely employed [17]. The inverted index is a word-oriented mechanism designed to index a text collection efficiently, enabling faster search operations. It consists of all distinct words in the text collection, each linked to a list of documents in which it appears. The full inverted index allows direct access

to matching terms and their positions within the documents. For multiple documents, a separate occurrence list is stored for each pair of term documents.

When a retrieval system receives a user query represented as terms, it searches for the keywords in the indexed documents. If the query matches the indexed terms, the system retrieves a list of relevant documents, ranks them based on relevance, and returns the top-ranked results to the user. Several free and open-source frameworks are available to support research in indexing, retrieval, and ranking, including Apache Solr, Elasticsearch, Terrier IR, and Pyserini.

Solr⁸ is an open-source enterprise search platform built on Apache Lucene and written in Java. It provides features such as full-text search, advanced indexing, and rich document handling. Similarly, Elasticsearch,⁹ also built on Apache Lucene and written in Java, differs by having a dual license, including the commercial Elastic license. It offers full-text search, advanced indexing, and client libraries that support various programming languages, including Python.

Terrier IR [173]¹⁰ is a modular open-source platform, also written in Java, developed by the IR Group at the School of Computing Science, University of Glasgow. It supports multiple indexing strategies and is compatible with standard TREC and CLEF test collections. Pyserini [144]¹¹ is a Python toolkit focused on reproducible IR research, supporting both sparse and dense retrieval through integrations with Anserini¹² (built on Apache Lucene) and Facebook's Faiss library [82]. It offers self-contained resources, including queries, relevance judgments, prebuilt indexes, and evaluation scripts, facilitating the easy reproduction of experiments on standard IR test collections.

The key difference between Solr, Elasticsearch, Terrier IR, and Pyserini lies in their focus and intended use cases. Solr and Elasticsearch are tailored to enterprise search applications, offering robust features for scalability and usability, but with limited support for research-oriented experimentation. In contrast, Terrier IR and Pyserini are designed for academic and research purposes, making them well-suited for investigating and developing IR algorithms and techniques. Their modular architecture enables researchers to streamline the reproducibility, integration, and customization of retrieval models, indexing strategies, and ranking functions. Additionally, Terrier supports Python through the *pyterrier* library [154],¹³ which facilitates data preprocessing, experimentation, and integration with other Python libraries.

2.1.7 Models

IR models provide a principled approach to estimating the relevance between a query and a set of documents. The queries and documents are first represented in a standard form, such as vectors or probability distributions, after which a ranking function assigns a relevance score to each document for the given query. Different IR tasks require models tailored to their specific characteristics. Broadly, IR models can be classified into three main categories: (i) traditional models,

⁸<https://solr.apache.org/>

⁹<https://www.elastic.co/elastic-stack/>

¹⁰<http://terrier.org/>

¹¹<https://github.com/castorini/pyserini>

¹²<https://github.com/castorini/anserini>

¹³<https://pyterrier.readthedocs.io/>

consisting of set-theoretic, algebraic, and probabilistic approaches [17]; (ii) feature-based models, which employ machine learning techniques such as learning-to-rank [147]; and (iii) neural IR models, which leverage deep learning architectures to improve retrieval effectiveness [161].

2.1.7.1 Classic IR Models

Classic IR models encompass a broad range of approaches, from exact matching Boolean to probabilistic models like BM25 [188] and language models [112, 247]. BM25 is a widely adopted probabilistic ranking function known for its strong performance in traditional IR baselines. Likewise, the probabilistic language model (LM) with Dirichlet smoothing [155] has proven effective, particularly for short queries in ad-hoc text retrieval tasks [247].

Moreover, the divergence from randomness (DFR) variant of BM25 (known as DFR BM25) has shown competitive results across various retrieval settings and has been validated in multiple TREC evaluations [176, 238]. Additionally, the Hiemstra LM [111, 112] has exhibited strong performance in ad-hoc text retrieval tasks, particularly in LRL scenarios [191]. These models serve as the foundation for the experimental design presented in this thesis. To facilitate the presentation of the models, we adopt a set of notations from Mitra and Craswell [161] and Lin et al. [145], which are defined as follows:

Let D denote the document collection and d a single document within that collection. Each term is denoted by t , and a query is represented as q , which consists of terms $t_q \in q$. The term frequency $tf(t_q, d)$ refers to the number of times the term t_q appears in document d , while $idf(t_q)$ denotes the inverse document frequency of t_q , and df is the document frequency. The $|d|$ refers to the length of d in words, $avgdL$ is the average document length in the collection, and $|D|$ is the total number of documents in the collection. The probability of a document d is denoted by $p(d)$, $p(t_q | D)$ is the probability of term t_q given the collection D , and $p(t_q | d)$ is the probability of term t_q given document d . The k_1 , b , c , μ , and λ are model-specific tuning parameters.

The traditional IR models are introduced in the following:

BM25 (*Okapi Best Matching*) BM25 [187, 188, 189] is a ranking function that estimates the relevance of a document for a given query based on the frequency of query terms in the document, while accounting for document length normalization and term frequency saturation. BM25 is defined as:

$$BM25(q, d) = \sum_{t_q \in q} idf(t_q) \cdot \frac{tf(t_q, d) \cdot (k_1 + 1)}{tf(t_q, d) + k_1 \cdot \left(1 - b + b \cdot \frac{|d|}{avgdL}\right)} \quad (2.1)$$

where k_1 is typically set between 1.2 and 2.0, and b is usually set to 0.75.

Term frequency (tf) [151] measures the frequency of a term within a document, while *inverse document frequency* (idf) [209] quantifies the importance of a term by considering the number of documents in the collection that contain that term.

The tf and idf are defined as follows:

$$tf(t, d) = f(t, d) \quad (2.2)$$

$$idf(t) = \log \frac{|D|}{df(t)} \quad (2.3)$$

To prevent division by zero when the term is absent, 1 is typically added to the denominator of idf .

One variant of BM25 is the **divergence from randomness model (DFR BM25)** [11, 238], which estimates the informativeness of a term by measuring the divergence between its frequency in a specific document and its frequency in the entire collection. It is formulated as:

$$weight(t|d) = \frac{tfn}{tfn + 1} \cdot \log_2 \left(\frac{|D| - df + 1}{df + 0.5} \right) \quad (2.4)$$

with:

$$tfn = tf \frac{t}{d} \cdot \log_2 \left(1 + c \cdot \frac{avgdl}{|d|} \right) \quad (2.5)$$

where tfn denotes normalized term frequency and c is typically set to 1.

Language Models (LMs) LMs apply statistical language modeling techniques to IR [147, 177]. According to Banerjee and Han [19], there are three typical approaches to applying LMs in ranking: (i) the query likelihood approach, which trains an LM on a document and estimates the probability of generating the query; (ii) the document likelihood approach, which trains a LM on the query and estimates the probability of generating the document; and (iii) the comparative approach (or model divergence), which trains language models on both the document and the query and then compares them.

In the language modeling-based approach, documents are ranked based on the posterior probability of a document given the query terms [111, 177, 247]. The formulation of this approach is summarized by Mitra and Craswell [161] as follows.

$$p(d|q) = \frac{p(t_q|d) \cdot p(d)}{\sum_{\vec{d} \in D} p(q|\vec{d}) \cdot p(\vec{d})} \quad (2.6)$$

$$\propto p(t_q|d) \cdot p(d) \quad (2.7)$$

$$= p(t_q|d), \text{ assuming } p(d) \text{ is uniform} \quad (2.8)$$

$$= \prod_{t_q \in q} p(t_q|d) \quad (2.9)$$

In its simplest form, the probability of a query term given a document can be estimated as:

$$p(t_q|d) = \frac{tf(t_q, d)}{|d|} \quad (2.10)$$

The application of language models to ad-hoc retrieval typically requires smoothing techniques to address data sparsity and the zero-frequency problem. These techniques incorporate term frequency information from both the document and the collection to produce more robust probability estimates [247]. Two commonly used smoothing techniques are Jelinek–Mercer smoothing and Dirichlet prior smoothing [161]. Zhai and Lafferty [247] observed that Jelinek–Mercer smoothing tends to perform better for long queries, while Dirichlet smoothing is generally more effective for short queries. The formulations of these two methods are described below:

1. **Jelinek–Mercer Smoothing** [17, 124] is a linear interpolation technique that combines the maximum likelihood estimate of a term based on document-specific frequency with its probability in the overall collection. This method balances document-specific evidence with collection-wide statistics to mitigate data sparsity. This method is formulated as:

$$p(t_q | d) = \lambda \cdot \frac{tf(t_q, d)}{|d|} + (1 - \lambda) \cdot \frac{\sum_{\bar{d} \in D} tf(t_q, \bar{d})}{\sum_{\bar{d} \in D} |\bar{d}|} \quad (2.11)$$

where $\lambda \in [0, 1]$ is the smoothing parameter. A higher value of λ tends to favor documents containing all query terms. Thus, careful tuning of λ is essential to improve performance.

2. **Dirichlet Prior Smoothing** [155, 247] is a Bayesian smoothing technique that models term frequencies using a multinomial distribution with a Dirichlet prior. The Dirichlet distribution, being a conjugate prior to the multinomial, enables efficient Bayesian inference within language modeling and is widely used in query likelihood models. It is defined as:

$$p(t_q | d) = \frac{tf(t_q, d) + \mu \cdot \frac{\sum_{\bar{d} \in D} tf(t_q, \bar{d})}{\sum_{\bar{d} \in D} |\bar{d}|}}{|d| + \mu} \quad (2.12)$$

where μ is the smoothing parameter. The document length $|d|$ is effectively augmented by μ , which controls the influence of the collection-wide term distribution in the smoothed probability estimate.

Moreover, **Hiemstra Language Model (LM)** [112, 186, 193] is another language modeling approach that incorporates smoothing for document ranking. It estimates the probability of a document given a query by computing the maximum likelihood of each query term appearing in the document. This is achieved through linear interpolation between the term’s relative frequency in the document and its relative frequency in the entire collection. The model is formulated as:

$$p(d, q) = p(d) \cdot \prod_{\langle i, t_q \rangle \in q} ((1 - \lambda_i) \cdot p(t_q | D) + \lambda_i \cdot p(t_q | d)) \quad (2.13)$$

where $\lambda \in [0, 1]$ is the smoothing parameter. When $\lambda = 0$, the model will produce random results, leading to near-zero average precision. As λ increases, retrieval performance typically improves, reaching an optimal point before gradually declining with further increases. Based on experimental evaluation, Hiemstra [111] recommended setting $\lambda = 0.15$ for optimal effectiveness.

2.1.7.2 Learning to Rank

Traditional ranking models, such as BM25, used predefined scoring functions based on statistical term-weighting schemes like term frequency and inverse document frequency. While these models have achieved strong performance across many IR tasks, they often struggle to generalize to new or unseen queries [147]. To address these limitations, learning to rank (LTR) approaches have been developed, which leverage machine learning techniques to learn ranking functions directly from data. A common approach within LTR is feature-based learning to rank, which relies on a set of manually engineered features extracted from query-document pairs. These features serve as input to a supervised learning algorithm trained to predict an optimal ranking of documents for a given query. Figure 2.2 illustrates the feature-based LTR framework.

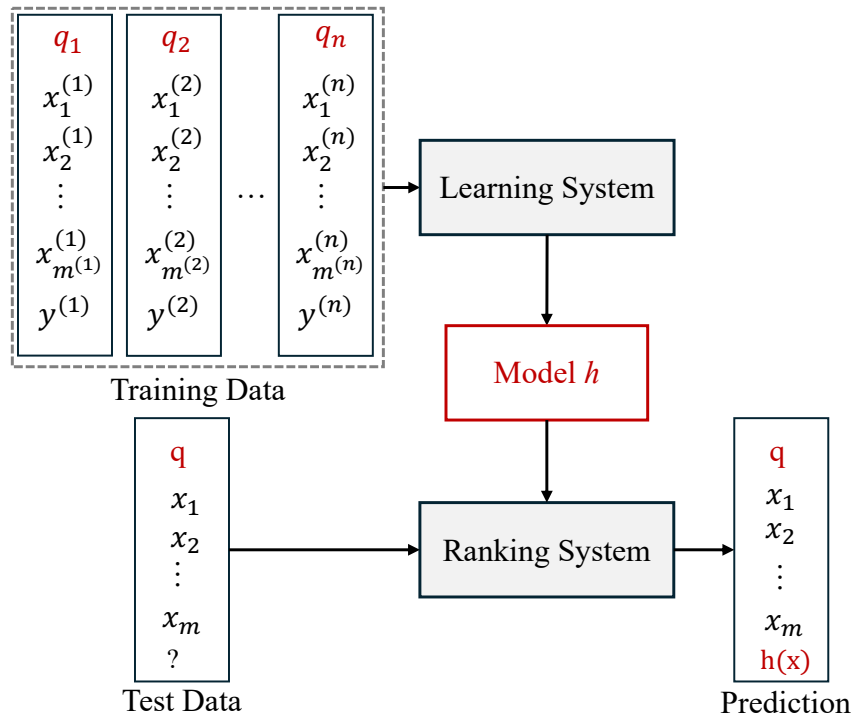


Figure 2.2: Learning to rank framework [147].

Feature-based LTR is a supervised learning approach that requires both training and test sets [35, 147]. A typical training set consists of n training queries q_i ($i = 1, \dots, n$), each associated with a set of candidate documents represented by feature vectors $X^{(i)} = x_j^{(i)}_{j=1}^{m^{(i)}}$, where $m^{(i)}$ denotes the number of documents linked to a query q_i . Each document is paired with its corresponding query and relevance label, indicating its degree of relevance. A learning algorithm is then applied to this labeled data to induce a ranking model, which is trained to minimize a predefined loss function by approximating the ground-truth relevance labels. At inference time, when a new query is provided, the trained model ranks its associated documents based on the predicted relevance scores and returns the resulting ranked list as the system's output.

There are three primary approaches in LTR [147]: pointwise, pairwise, and listwise. Pointwise

methods treat ranking as a regression or classification problem by assigning a relevance score or label to each document. The model learns to predict the relevance of a single document given a query, independent of other documents. Pairwise approaches, by contrast, model the relative ordering between pairs of documents. The learning objective is to predict which document in a given pair should be ranked higher, aiming to minimize the number of incorrectly ordered pairs. Listwise approaches consider the entire set of documents associated with a query and directly optimize for ranking metrics, such as Mean Average Precision (MAP) or Normalized Discounted Cumulative Gain (NDCG). The evaluation metrics are outlined in Subsection 2.1.9.

In text retrieval, the effectiveness of pointwise, pairwise, and listwise approaches has been extensively studied, with listwise methods generally achieving the best performance [35, 147, 180]. This advantage is mainly attributed to their ability to optimize ranking quality across the entire set of documents associated with a query. By considering the global structure of the ranked list, listwise methods better capture both document relevance and positional dependencies, resulting in more effective rankings.

2.1.7.3 Neural IR Models

In contrast to LTR models that rely on manually engineered features, neural IR models automatically learn relevance signals by mapping queries and documents into low-dimensional vector spaces (or *embeddings*). At a high-level, document ranking in neural IR typically involves three main steps [161]: (1) generating a query representation that captures the user’s information need, (2) generating a document representation that reflects the semantic content of the document, and (3) matching these representations to estimate the relevance of each document to the given query.

Neural approaches to IR are generally categorized based on their impact on query representation, document representation, or the relevance estimation process. The general framework for neural IR in document ranking is illustrated in Figure 2.3.

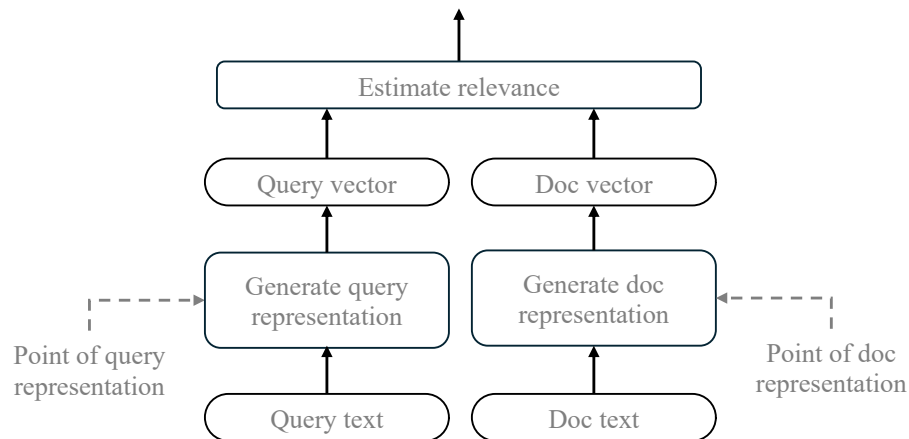


Figure 2.3: General neural information retrieval approach for document ranking [161].

The application of neural models to ad-hoc document retrieval was first explored by Yang et al. [243], who applied Bidirectional Encoder Representations from Transformers (BERT) [79].

BERT, built on the transformer architecture [227], is particularly effective at capturing contextual relationships through its dual encoders. In their approach, inference was performed at the sentence level, and the resulting relevance scores were aggregated to estimate overall document relevance. Building on this line of work, Nogueira et al. [167] fine-tuned the pre-trained Text-to-Text Transfer Transformer (T5) [184], demonstrating that it outperformed BM25 with only 20k training examples on the MS MARCO passage dataset [18]. These early breakthroughs laid the groundwork for a rapidly expanding body of research aimed at refining neural approaches for ad-hoc text retrieval [145, 234, 252].

Despite these advancements, training and fine-tuning neural retrieval models (dense retrieval models) remain computationally expensive and typically require large-scale labeled datasets, which are often scarce or costly to obtain for IR tasks [129]. As an alternative, zero-shot text retrieval has gained significant attention in recent years [39, 129, 185, 216, 249]. More recently, researchers have investigated the potential of large language models (LLMs) to support retrieval tasks by leveraging their pretrained knowledge to enhance performance [123, 255].

2.1.8 Test Collections

The standard method for creating test collections is based on the Cranfield paradigm, introduced in the 1960s and later popularized through the TREC series of large-scale evaluation campaigns that began in 1992. The Cranfield approach for comparative evaluation involves assessing the performance of various retrieval strategies or systems. This process typically includes the following stages [42]: (1) Selecting different retrieval strategies or systems to compare and generating top-ranked lists of documents (*runs*) for each query (*topic*); (2) Compute the effectiveness of each strategy for every query in the test collection as a function of relevant documents retrieved; (3) Averaging the scores across all queries to determine the overall effectiveness of each strategy; and (4) Use the score to rank the strategies based on their relative performance to each other.

In the TREC-style implementation of the Cranfield approach, the *runs* submitted by participating systems are pooled to create a set of documents for each query. Human annotators then assess the relevance of these documents, producing query-document relevance judgments (*qrels*). A standard TREC-style test collection typically consists of three components: a document corpus, a set of information needs (*topics*), and the corresponding set of relevance judgments (*qrels*) [107]. TREC includes many tracks spanning a variety of test collections, but the best known are those from the ad-hoc track during the first eight TREC evaluations, held between 1992 and 1999 [158, 231]. The following subsections describe the common components of test collections used in the ad-hoc track across TREC, NTCIR, CLEF, and FIRE.

2.1.8.1 Documents

In text retrieval, documents are a collection of structured, semi-structured, or unstructured natural language texts. A test collection must contain a static set of documents to ensure that evaluation

results are reproducible when reused. Each document is typically tagged using Standard Generalized Markup Language (SGML) or Extensible Markup Language (XML) and assigned a unique identifier, *DOCNO*. A typical document consists of an identifier, a title, and a body of text. In the ad-hoc TREC track, a document collection is designed to represent a diverse range of topics [233].

2.1.8.2 Topics

In search, *topics*, commonly referred to as queries, are short texts entered by users into a search box to express their information needs. In the ad-hoc retrieval tasks, a topic typically consists of: (i) *Identification*—a unique identifier for the topic; (ii) *Title*—a brief description of the information needed, typically in the form of a short text; (iii) *Description*—a sentence providing additional context about the desired information; (iv) *Narrative*—a detailed paragraph describing the characteristics of the documents expected to satisfy the information need.

2.1.8.3 Relevance Judgments

Relevance judgment is the process of evaluating the relevance of documents to a given query based on human assessment. These judgments are typically categorized into binary relevance or graded relevance. Binary relevance classifies each document as either relevant or non-relevant to the query, assigning a score of 1 to relevant documents and 0 to non-relevant ones. Graded relevance assesses documents on multiple levels of relevance, with higher scores indicating a greater degree of relevance to the information need.

Binary relevance has been widely used in experimental research, particularly within TREC collections. In the TREC-9 Web Track [108], three-level graded relevance was introduced, categorizing documents as not relevant, relevant, or highly relevant. Later, Sormunen [208] proposed a four-level relevance scale consisting of non-relevant, marginally relevant, relevant, and highly relevant. Kekäläinen [127] refined this four-point scale, renaming the third level to “fairly relevant” instead of “relevant”. This updated four-point scale has been adopted in various TREC tracks, including the ad-hoc track.

2.1.9 Evaluation Metrics

Evaluation metrics measure the effectiveness of IR systems in retrieving documents relevant to a user’s query. To standardize IR evaluation, Clough and Sanderson [42] categorized these metrics into set-based, rank-based, and other measures. The notation used in the following formulas is based on that introduced by Mitra and Craswell [161].

2.1.9.1 Set-Based Measures

Set-based measures evaluate retrieval effectiveness without considering the order of the ranked results. Two commonly used metrics in this category are precision and recall.

Precision Measures the fraction of retrieved documents that are relevant for a given query q relative to the total number of documents in the retrieved set R_q . It is defined as:

$$Precision_q = \frac{\sum_{\langle i,d \rangle \in R_q} \text{rel}_q(d)}{|R_q|} \quad (2.14)$$

where $\langle i,d \rangle$ denotes a tuple corresponding to document d at rank i position, with $\langle i,d \rangle$ in R_q . The function $\text{rel}_q(d)$ is an indicator that returns 1 if document d is relevant to query q and 0 otherwise.

Recall Calculates the fraction of relevant documents retrieved for a query q relative to the total number of relevant documents in the collection D . It is formulated as:

$$Recall_q = \frac{\sum_{\langle i,d \rangle \in R_q} \text{rel}_q(d)}{\sum_{d \in D} \text{rel}_q(d)} \quad (2.15)$$

2.1.9.2 Rank-Based Measures

Rank-based measure evaluation focuses on the order of the ranked retrieval results. A common approach is to calculate the precision at a fixed rank position, such as the precision at rank 10 (P@10), which computes the proportion of relevant documents among the top 10 documents retrieved. For multiple queries, precision scores are averaged to calculate the mean average precision (MAP). For non-binary relevance judgments, the normalized discounted cumulative gain (NDCG) is used to account for graded relevance levels.

Mean Average Precision (MAP) This metric computes the mean of the average precision [254] scores across all queries Q . The average precision (AP) combines precision and recall across ranked retrieval results, providing a single metric that reflects the ranking quality by considering both the relevance of documents and their positions in the ranked list. The AP is defined as:

$$AP_q = \frac{\sum_{\langle i,d \rangle \in R_q} Precision_{(q,i)} \times \text{rel}_q(d)}{\sum_{d \in D} \text{rel}_q(d)} \quad (2.16)$$

where $Precision_{q,i}$ represents the precision calculated at rank i for the query q .

The mean average precision (MAP) for all queries Q is then formulated as:

$$MAP = \frac{1}{Q} \sum_{q \in Q} AP_q \quad (2.17)$$

Normalized Discounted Cumulative Gain (NDCG) Represents a normalized version of discounted cumulative gain (DCG), which evaluates the relevance of documents in a ranked list based on a graded relevance scale and their positions [121]. The DCG penalizes highly relevant documents that appear lower in the ranking by reducing their relevance score logarithmically to their position. The DCG for a query q accumulated at rank i is defined as:

$$DCG_q = \sum_{\langle i,d \rangle \in R_q} \frac{2^{rel_q(d)} - 1}{\log_2(i + 1)} \quad (2.18)$$

where $rel_q(d)$ is the graded relevance score of document d at position i for the query q .

To account for variations in the length of result lists across queries or systems, NDCG normalizes DCG using the ideal discounted cumulative gain (IDCG). The IDCG represents the DCG of an ideal ranking, where documents are ordered by their relevance. NDCG is formulated as:

$$NDCG_q = \frac{DCG_q}{IDCG_q} \quad (2.19)$$

where $IDCG_q$ is the DCG of the ideally ranked result list for query q .

2.1.9.3 Other Measures

Mean Reciprocal Rank (MRR) [50, 51] is a commonly used metric for evaluating the performance of question-answering and search systems. It measures the effectiveness of systems that return a ranked list of responses by considering the position of the first relevant result.

The reciprocal rank (RR) for a single query q is defined as:

$$RR_q = \max_{\langle i,d \rangle \in R_q} \frac{rel_q(d)}{i} \quad (2.20)$$

where R_q is the ranked list of retrieved documents for query q , i is the rank position, and $rel_q(d)$ is an indicator function that equals 1 if document d is relevant to query q , and 0 otherwise.

MRR is then computed as the average reciprocal rank across all queries Q :

$$MRR = \frac{1}{|Q|} \sum_{q \in Q} RR_q \quad (2.21)$$

2.1.10 Notion of Relevance

Relevance is the foundation of all ranking metrics in IR, serving as a fundamental criterion for evaluating how effectively a retrieval system addresses user search needs. Relevance can take various forms, including content (topic, quality, etc.), object, validity, situational, cognitive, affective, or belief relevance [197]. In evaluating search performance, IR researchers typically focus on *topical* relevance [145], which reflects the degree to which a retrieved document matches the user's information need based on its topical content.

A document is considered relevant if it satisfies the user's information need expressed through a natural language text query. A key challenge in assessing relevance lies in its inherently subjective nature, which is influenced by contextual factors such as the user's task, time, location, or device [17]. Given that no IR system can provide perfect answers to all users at all times, diverse retrieval strategies must be explored and evaluated to effectively address the complexity and variability of user information needs.

2.2 Tetun

Tetun, also written as “Tetum” in English or “Tétum” in Portuguese, is an Austronesian language spoken in Timor-Leste, an island nation in Southeast Asia located between Indonesia and Australia (see [Figure 2.4](#)). Timor-Leste first declared independence from Portugal on November 28, 1975, but was invaded by Indonesia just over a week later. Independence was officially restored on May 20, 2002, making it the first new sovereign state of the 21st century.



Figure 2.4: Map of Timor-Leste (source: <https://www.friendsofossu.org.au/about-ossu-timor-leste/>).

There are two main varieties of Tetun: Tetun Dili, also known as Tetun *Prasa* (commonly referred to simply as Tetun), and Tetun Terik [224]. Tetun has two standardized forms: one produced by the *Instituto Nacional de Linguística* (Tetun INL) and the other by the Dili Institute of Technology (Tetun DIT), while Tetun Terik is a dialect spoken in Timor-Leste.

Tetun is one of Timor-Leste’s official languages, alongside Portuguese [226]. The government established Tetun INL as the official standard orthography, which is used in the education system, official publications, and media [169]. Tetun DIT, developed by linguists at the Dili Institute of Technology, differs from Tetun INL in certain writing conventions [224]. For example, the words [*fó* (give), *ne’ebé* (which/that)] in Tetun INL correspond to [*foo*, *neebe*] in Tetun DIT.

According to the 2015 census, Timor-Leste had a population of 1.18 million, with 78.78% identified as Tetun speakers [57, 62]. Of these, 30.50% reported Tetun as their home language, while 48.28% spoke it as a second or third language [57, 60]. The 2022 census recorded a population growth of 13.40%, with the total increasing from 1.18 million to 1.34 million [58], but it did not include updated statistics on Tetun speakers.

2.2.1 Evolution

The evolution of Tetun is illustrated in [Figure 2.5](#). The first known Tetun materials appeared at the end of the 19th century in the Catholic catechism written by a Portuguese priest, Sebastião Maria

Aparício da Silva [103, 223], in the era of Portuguese colonialism in Timor-Leste, which lasted from 1702 to early October 1975 [104].

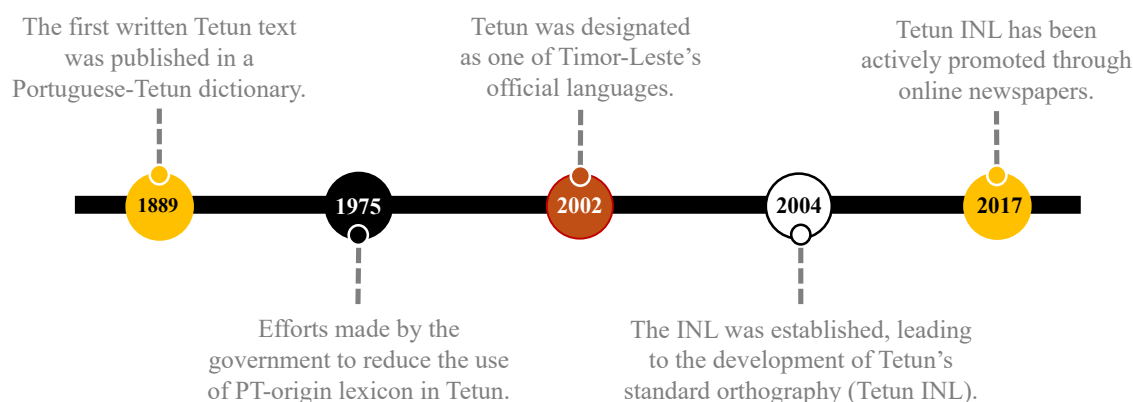


Figure 2.5: The evolution of Tetun.

Throughout this period, Portuguese authors produced works in Tetun, and consequently, Portuguese orthographic rules were directly applied to Timorese Tetun [103]. As a result, the first Tetun vocabulary in Silva's catechism included the cedilla (ç), and acute and circumflex accents. For example, *oraçã* ("orasaun" in Tetun, meaning "prayer"), *boçoc* ("bosok" in Tetun, meaning "lie"), and *cátac* ("katak" in Tetun, meaning "that").

In November 1975, Timor-Leste declared its independence, and the FRETILIN (*Frente Revolucionária de Timor-Leste Independente*) attempted to consolidate the orthography of Tetun and promote it as a national language through the publication of *Como Vamos Alfabetizar o Nosso Povo Mau Bere de Timor-Leste* (How Can We Provide Literacy to Our Mau Bere People of Timor-Leste) [103]. This work aimed to improve the phonemic representation of Tetun and reduce Portuguese-based orthography, such as replacing the circumflex accent in "-ão" with "-aun".

However, this consolidation process was interrupted by Indonesia's invasion in December 1975, after which Timor-Leste was declared the 27th province of Indonesia. During the occupation, Tetun was used mainly in trade and religious activities until Timor-Leste voted for independence in early September 1999.

After independence was formally restored on May 20, 2002, the government designated Tetun as one of the country's official languages, alongside Portuguese [226]. Since then, it has become a dominant language in public life. In 2004, the government established the *Instituto Nacional de Linguística* (INL, National Institute of Linguistics) and produced the standard Tetun orthography, known as "Tetun INL", which is now the official form used in the education system, official publications, and media [169].

2.2.2 Evolution in Newspapers

After Timor-Leste voted for independence in 1999, Tetun began to appear more frequently in newspapers [223]. The United Nations published the first printed newspaper in Tetun in October 1999. In July 2000, *Voz de/Suara Timor Lorosa'e* (STL) became the first commercial news agency

to publish in Tetun alongside Indonesian. By 2017, more than five print newspapers were regularly publishing news in Tetun [45].

The first online news outlet in Tetun appeared in 2008, launched by *Tempo Semanal*.¹⁴ However, after five years, the portal stopped updating in January 2013 and remains inactive, although the print edition continues to circulate. In March 2017, the government of Timor-Leste launched Tatoli [170],¹⁵ which significantly boosted the use of Tetun, particularly the Tetun INL standard.

Two years later, in May 2019, Borulili Timor News launched Timor News [59], an online portal dedicated exclusively to national and international news in Tetun. On October 30, 2020, Timor News was officially registered with the Press Council of Timor-Leste¹⁶ as a news agency on October 30, 2020, under the sole ownership of the author of this thesis [56]. By 2021, more than ten online news agencies were actively publishing daily articles in Tetun.

2.2.3 Orthography

Tetun INL uses the Latin script, consisting of 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, ', v, x, z* [168]. The letters *c, q, w*, and *y* are not used in Tetun except in proper names and international symbols.

Accented vowels *á, é, í, ó, ú* are also used, and the apostrophe (*'*) denotes a glottal stop. Additionally, the hyphen is employed to indicate monosemantic compound words. Some basic Tetun phrases are presented in Figure 2.6.

Tetun	English
Dadeer di'ak!	Good morning!
Di'ak ka la'e?	How are you?
Ita-nia naran saida?	What is your name?

Figure 2.6: Examples of basic Tetun phrases.

2.2.4 Morphology

Morphology is conventionally divided into inflection and word formation, with word formation further categorized into derivation and compounding [15]. Inflection refers to the different syntactic variations of a word that do not alter its core meaning, while word formation involves the creation of new nouns, verbs, and adjectives. Derivation creates a new word from an existing one, whereas compounding combines two or more words to form a new word.

Morphological processes such as circumfixes and reduplication also contribute to both the formation of new words and the modification of existing word structures. Circumfixes involve the

¹⁴<http://temposemanaltimor.blogspot.com/>

¹⁵<https://tatoli.tl>

¹⁶<https://conselhoimprensa.tl>

simultaneous addition of a prefix and a suffix to a base word, while reduplication is a morphological process in which a part of a word is copied, either fully or partially, to form a new word that may have additional morphemes attached to it [103, 222].

Tetun does not have rich inflectional and derivational morphology, with only a few inflectional affixes [103, 116, 224]. Tetun affixes include both native Tetun elements and those derived from Portuguese. Prefixes are exclusive of native Tetun, whereas suffixes can derive from either native Tetun or Portuguese. In compounding, words are combined using hyphens, exclusively with native Tetun. Additionally, Tetun uses circumfixes and reduplication within its native vocabulary and adopts zero derivation for Portuguese-derived words. Examples of Tetun inflection and derivation are illustrated in Figure 2.7.

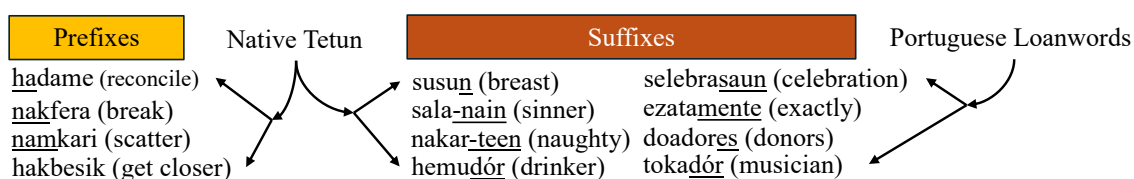


Figure 2.7: Examples of Tetun inflection and derivation. The suffix *dór* is used in both native Tetun and Portuguese loanwords.

Circumfixes and reduplication are not widely used in Tetun. The circumfixes are not productive [103] and are confined to simple verbs, typically consisting of one or two syllables derived from verbs [116]. Reduplication is similarly limited, being applied only to nouns, adjectives, adverbs, and numerals, with a few instances of its use for pluralization [103].

2.2.5 Loanwords

A significant portion of Tetun verbs, nouns, and adjectives are derived from Portuguese, an influence especially evident in the news media [103, 106, 223, 224]. Examples of Portuguese loanwords are shown in Figure 2.8.

“Ha’u **agradese** **tebes**
tanba halo **investigasaun**
ida-ne’ebé **krusiál**.”

Verb Noun Adjective

Figure 2.8: Examples of Portuguese loanwords: *agradese* (*agradecer*, meaning “thank”), *investigasaun* (*investigação*, meaning “research”), *krusiál* (*crucial*, meaning “crucial”).

Klinken et al. [224] reported that the prevalence of Portuguese loanwords can be traced back to Portuguese-educated political leaders, many of whom continued using Portuguese at home after 1975. Following independence, the frequent appearances of these leaders in the news media further contributed to the incorporation of Portuguese vocabulary into spoken and written Tetun.

Klinken and Hajek [223] analyzed seven newspaper articles from 2009 and found that, on average, 32% of words were Portuguese loanwords. Similarly, Greksáková [103] reported that 35% of the vocabulary in a corpus of 73,892-word interview transcripts consisted of Portuguese loanwords. Additionally, Hajek and van Klinken [106] noted that Portuguese influence in Tetun, particularly in newspapers and technical documents, had increased to over 40%, with headlines often composed almost entirely in Portuguese.

In a recent study, de Jesus and Nunes [62] found that 28.20% of words in Tetun were Portuguese loanwords, based on an analysis of approximately 10.69 million words from the Labadain-30k+ dataset [64], covering the period from 2017 to 2023. Additionally, they also observed a 5.09 percentage-point increase in Portuguese loanword adoption when comparing documents created before and after 2017.

2.2.6 Text Documents

As of 2023, fewer than 45,000 Tetun text documents are available on the web [62, 136]. Kudugunta et al. [136] introduced the MADLAD-400 dataset, a multilingual corpus created by Google Research and Google DeepMind teams, which includes 40.4k Tetun documents. Released in October 2023, this dataset was compiled from CommonCrawl snapshots covering 2008 to August 2022.

More recently, de Jesus and Nunes [64] released the Labadain-30k+ dataset, a monolingual Tetun collection comprising 33.6k documents spanning from 2010 to 2023. This dataset was constructed through web crawling and enriched with Tetun Wikipedia documents published by the Wikimedia Foundation [171].

The main differences between these two datasets lie in document quantity and structure. The Labadain-30k+ dataset contains approximately 6.8k fewer documents than the Tetun documents in the MADLAD-400 dataset. However, it provides richer contextual metadata, including URLs, document sources, publication dates, content categories, and full text. In contrast, the Tetun documents in MADLAD-400 are limited to plain text without associated metadata.

Chapter 3

Text Information Retrieval in Low-Resource Languages

This chapter discusses text information retrieval (IR) in the context of low-resource languages (LRLs), focusing on aspects relevant to this thesis. It begins with an overview of key characteristics of LRLs, followed by techniques for dataset construction and the development of algorithms and preprocessing methods tailored to these languages. Finally, it highlights retrieval and ranking experiments that align with this thesis’s objectives.

3.1 Overview

Low-resource languages, also referred to as under-resourced or less-resourced languages [25, 40], are characterized by low data density (i.e., limited availability of online resources), a lack of textual corpora, reduced accessibility, and limited computational resources (such as annotated datasets, algorithms, and tools) [25, 40, 156, 200]. These languages are less studied in natural language processing (NLP) and IR research and remain underserved in the broader online ecosystem.

Tetun falls into this category due to several limiting factors, including data scarcity (i.e., a limited number of publicly available web documents [136, 62]), the lack of established computational resources, and the absence of NLP- and IR-related research before this thesis.

3.2 Dataset Construction Pipelines

Text corpora are fundamental for advancing language technology tools, particularly in the IR and NLP domains. In recent years, constructing such corpora for LRLs has gained significant attention [6, 109, 137, 156]. Despite the scarcity of resources, there has been a growing interest in addressing the needs of these languages [26]. Although various techniques have been developed to mitigate the challenges of constructing datasets for LRLs, several persistent issues remain unresolved. These include the lack of high-quality datasets, the absence of standardized writing rules, and the informal nature of many written texts [132, 134, 146]. Other challenges

include data scarcity, the lack of annotated data, and the limited availability of Wikipedia documents [200, 211, 212, 244].

To overcome these challenges, researchers have proposed a variety of methods, ranging from the use of tools like the BootCaT toolkit [20] to more recent approaches such as manually scraping high-quality documents from websites [16], employing collaborative and crawling methods [133], extracting data from Common Crawl [235], modifying existing frameworks such as Nutch for crawling [215], or building a data collection pipeline from scratch [146]. Research has also focused on addressing challenges specific to LRL corpus construction, including improving the quality of web-crawled datasets and handling language-specific issues [1, 6, 134]. The following subsections provide detailed descriptions of these techniques and tools.

3.2.1 BootCaT Toolkit

Baroni and Bernardini [20] introduced the BootCaT toolkit,¹ which involves manually selecting a list of 5 to 15 seed words from a specialized domain. These seed words are used as queries via the Google API to retrieve the top- n documents, which are then formatted as text.

The retrieved documents are subsequently analyzed to generate new seed words by randomly selecting terms from the corpus to create additional queries in an iterative process that can be repeated as needed. Initially tested in English and Italian, this approach has been extended and applied in various studies, including the construction of LRL datasets through the Crúbadán project [200], such as the Irish dataset, which comprises 125,000 documents.

3.2.2 Tailored Crawling

Artetxe et al. [16] proposed a tailored crawling that involved manually identifying data sources containing high-quality content and subsequently scraping their content to construct a text corpus. This technique was applied to Basque, resulting in the creation of EusCrawl, a Basque text corpus comprising 12.5 million documents and 423 million tokens. The quality of EusCrawl was manually audited and compared with the Basque content extracted from mC4 [242] and CC-100 [44]. The evaluation process involved randomly sampling 100 documents from each corpus (300 documents in total), which were then assessed by native Basque-speaking annotators.

The quality assessment of the documents considered several factors: *language* (whether the document is in Basque), *language variety* (whether the document is written in standard and correct Basque), *coherence* (whether the document contains gaps or disconnected portions), *noise* (whether the document is clean and free of extraneous elements), and *content* (whether the document appears to be automatically generated or lacks meaningful information). Annotators also classified the overall quality of each document as high, medium, or low. The qualitative evaluation demonstrated that tailored crawling is an effective approach for constructing high-quality datasets.

¹<https://bootcat.dipintra.it>

3.2.3 CURL Pipeline

Körner et al. [133] introduced Crawling Under-Resourced Languages (CURL), a data collection pipeline designed to facilitate community participation in corpus creation through a publicly accessible web portal.² To submit a new corpus, users select a language from the available list and either provide a list of URLs or upload plain text files for processing. CURL employs the Heritrix crawler [162]³ to crawl the specified URLs and uses jWarcEx⁴ to extract text content from HTML documents. The extracted text is then filtered with Brown’s LID model [31] to exclude non-target language documents.

The retrieved documents also underwent preprocessing steps, including sentence segmentation, rule-based cleaning, language separation into individual sentences, and sentence deduplication. Through its implementation, CURL has facilitated the creation of corpora for 258 LRLs, with Iranian Persian achieving the largest corpus size, comprising over 3.9 million sentences.

3.2.4 Corpulyzer Framework

Tahir and Mehmood [215] proposed Corpulyzer, a framework for corpus generation and analysis tailored to creating corpora for LRLs. Built by modifying the Apache Nutch framework, this approach was applied to Urdu by first filtering an existing Urdu corpus extracted from Common Crawl. This filtering process utilized the Compact Language Detector 2 (CLD2)⁵ to generate a list of seed URLs. Corpulyzer was then employed to crawl the web, resulting in the creation of UrduWeb20, a dataset containing 4.1 billion tokens extracted from eight million web pages and 6,590 unique URLs.

3.2.5 Common Crawl Processing Pipeline

Wenzek et al. [235] presented a pipeline for processing Common Crawl (CC) snapshots to construct text corpora. The process involves downloading each CC snapshot and processing it independently to extract language-specific text data using the fastText language identification (LID) classifier [102]. To ensure data quality, each page in the dataset is tokenized using the SentencePiece tokenizer [135], and text perplexity is calculated using a language model trained on a targeted domain (e.g., Wikipedia) as a quality metric.

This pipeline resulted in the creation of a monolingual text dataset containing 3.2 TB of compressed documents across 174 languages, including several LRLs. Notable examples include Basque, with approximately 10.36M sentences, and Malay, comprising roughly 6.96M sentences.

²<https://curl.wortschatz-leipzig.de/>

³<https://github.com/internetarchive/heritrix3>

⁴<https://github.com/Leipzig-Corpora-Collection/jwarcex>

⁵<https://github.com/CLD2Owners/cld2>

3.2.6 SwissCrawl Pipeline

Linder et al. [146] introduced SwissCrawl, a crawling tool developed from scratch to collect web documents and construct a corpus for Swiss German (GSW). The system has two main components: a *seeder*, which identifies potential URLs using a search engine, and a *crawler*, which extracts GSW sentences from web pages.

The seeder begins by extracting a list of sentences from the Leipzig Web Corpus 2017 [101] and filtering them using a GSW LID model developed by the authors. Next, it computes the vocabulary frequency, normalizing words to lowercase and discarding non-alphabetic tokens. Words occurring only once or that are not identified as GSW are excluded. Query seeds are then generated by randomly sampling three words, with selection probabilities based on their frequency distribution. Seeds with a GSW LID score below 95% are discarded. Each valid seed is enclosed in double quotes and submitted to startpage.com, retrieving the first 20 unseen URLs, which are saved for further crawling.

The crawler processes the collected URLs and metadata, which are stored in a file or MongoDB instance, and adds them to a task queue initialized with zero depth. For each task, the content is scraped using [requests](https://requests.readthedocs.io/)⁶ and BeautifulSoup,⁷ with boilerplate removal handled by [jusText](https://ftfy.readthedocs.io/).⁸ The content is normalized using the [ftfy](https://ftfy.readthedocs.io/) library⁹ to fix Unicode issues and split into individual sentences using Moses' `split-sentences.perl`.¹⁰

Filtering is applied at multiple stages. Sentences are filtered based on predefined rules, and the GSW LID model is used to exclude non-GSW text. Outgoing URLs are filtered, and duplicate sentences are removed. The final corpus comprised 562,524 sentences collected from 62,000 URLs across 3,472 domains.

3.2.7 Summary

The BootCaT Toolkit provides an alternative solution for constructing datasets for LRLs. However, it faces challenges, particularly its reliance on manually selected seed words, which can introduce subjectivity, bias, and inefficiencies, especially for less experienced users. Tailored crawling is another effective technique for gathering large volumes of high-quality data, but its success depends on the availability of data sources with sufficient relevant content. This method is particularly useful for LRLs with readily accessible high-quality data.

The CURL project combines crawling with collaborative techniques, offering a robust solution for long-term data collection efforts. However, the major limitation lies in motivating the community to contribute content. Holy et al. [228] addressed this issue by employing gamification strategies with incentives to encourage community participation. While effective, this approach requires funding to sustain the reward system.

⁶<https://requests.readthedocs.io/>

⁷<https://www.crummy.com/software/BeautifulSoup/bs4/doc/>

⁸<http://hdl.handle.net/11858/00-097C-0000-000D-F696-9>

⁹<https://ftfy.readthedocs.io/>

¹⁰<https://github.com/moses-smt/mosesdecoder>

Using publicly available snapshots from Common Crawl to construct datasets introduces computational constraints, as processing these snapshots demands powerful computing resources. Similarly, SwissCrawl has limitations, including limited documentation and resources, while the Corpulyzer framework is not publicly accessible, which limits its wider adoption and application.

To overcome these challenges, we developed the Labadain Crawler [74], a data collection pipeline designed to address the challenges of constructing text corpora for LRLs. Its effectiveness was tested through its application to Tetun, resulting in the construction of a text corpus suitable for a wide range of IR and NLP tasks. Details on this solution are provided in Chapter 4.

3.3 Text Preprocessing Techniques

Text preprocessing involves transforming text from various formats into a standardized representation to improve retrieval effectiveness. It also includes filtering documents in the target language from web collections. The following subsections describe commonly used text preprocessing techniques, particularly in LRL scenarios.

3.3.1 Language Identification

Language identification (LID) is a critical step in processing web-crawled data for constructing linguistic corpora in LRLs [122]. King and Abney [131] proposed a word-based LID method for mixed-language text documents in LRLs using weakly supervised techniques. Their approach combined 1- to character 5-grams and full-word features, employing a logistic regression classifier. The dataset included small monolingual samples from 30 languages, with most text sizes ranging between 1,300 and 1,600 words. These texts were collected from the web using the BootCaT toolkit [20] and underwent manual evaluation by human annotators to create the final dataset. Four classifiers were tested: logistic regression (LR), Naive Bayes (NB), decision tree, and Winnow2. By varying the number of sampled words in each training example from 10 to 1,000, the Naive Bayes classifier achieved the highest accuracy in most experiments, with scores exceeding 0.7.

Espichán-Linares and Oncevay-Marcos [86] developed an LID model for the languages spoken in Peru, creating a text corpus for 49 indigenous languages and dialects by collecting documents from the web and private repositories. The preprocessing stage included manual correction and annotation, followed by a split into training and testing sets (70%–30%). Their experiments used character-based features, including bigrams, trigrams, and four-grams. The training set was used to evaluate five classifiers with 5-fold cross-validation, including Support Vector Machines (SVM) with a linear kernel and Multinomial Naive Bayes (MNB). The SVM classifier with a linear kernel achieved the best performance, with a precision of 98.70% during cross-validation and an average precision of 95.9% on the test set.

Jauhiainen et al. [122], in their survey on automatic LID in text, reported that traditional machine learning models—such as MNB, SVM, and LR—trained on 5-gram characters demonstrated competitive performance across various experiments involving diverse languages, including LRLs,

for both monolingual and multilingual text data. For closely related languages in multilingual settings, Zampieri [246] demonstrated that MNB outperformed SVM in training LID models, particularly when classifying similar languages such as European and Brazilian Portuguese.

3.3.2 Stopword Detection

Stopword detection has been extensively studied for high-resource languages, with methods that can also be adapted for LRLs. Fox [94] introduced the term frequency (TF) approach, which was initially applied to the Brown Corpus to develop a stopwords list for English. This method identifies the top- n most frequent terms in a corpus, followed by human validation to refine the list by removing non-stopwords. The TF approach has been widely used to create stopwords lists for LRLs, including Marathi and Czech [100], as well as Kinyarwanda and Kirundi [166].

Moreover, Lo et al. [148] proposed the normalized inverse document frequency (NIDF) method for stopwords detection. This method was tested on four English TREC collections and compared to TF, inverse document frequency (IDF), and term frequency-inverse document frequency (TF-IDF). The results showed that NIDF was more effective than TF and IDF in detecting stopwords.

More recently, Ferilli [91] introduced term-document frequency (TDF) and evaluated its effectiveness using corpora in Italian and Squinzanese (a dialect from southern Italy). The study demonstrated that TDF outperformed TF, IDF, and NIDF, particularly when applied to smaller corpora. TDF is calculated by multiplying the total number of a term's occurrences in the corpus by the total number of documents in which the term appears.

3.3.3 Stemming

Stemming is an NLP technique used to reduce words to their root form, known as a stem. While the stem may not always be a valid word in the language, it is intended to capture the core meaning shared by all related forms of the word. Stemming is widely employed in IR to group similar words, aiming to improve retrieval efficiency and effectiveness. Numerous stemming algorithms exist, but the Porter Stemmer [178] remains one of the most commonly used rule-based methods for suffix removal in English. This approach has also been adapted for several LRLs, including Czech [80], Sanskrit [191], Indonesian [2], and Sundanese [13].

The Czech stemmer [80] consists of two variants: *light* and *aggressive* stemmers. These approaches focus on removing suffixes from nouns and adjectives while excluding verbs, which are more prone to stemming errors. The light stemmer primarily removes inflectional suffixes, conflating singular and plural forms or feminine and masculine variants to a root stem. The aggressive stemmer extends this process by also removing certain frequent derivational suffixes. Both variants incorporate rules to address orthographic irregularities.

The Sanskrit stemmer [191] adopts a similar structure, also featuring *light* and *aggressive* variants. However, the aggressive variant is specifically designed to remove derivational suffixes. The suffix list for Sanskrit was developed by linguists through observations of word inflections in nouns, adjectives, and genders.

The Indonesian stemmer [2], known as confix-stripping, processes both suffixes and prefixes, unlike many stemmers that focus solely on suffixes. This approach combines a set of rules with root word dictionary lookups. A minimum word length of four characters is enforced, meaning words with three or fewer characters are excluded from processing. The Sundanese stemmer [13], built on the Indonesian stemmer, introduces modifications to the stemming sequences and incorporates handling for infixes and allomorphs (affixes formed by altering the initial phoneme of a stem to create a verb).

An advanced version of the Porter Stemmer is the Snowball stemming algorithm [204], which supports multiple languages, including Portuguese, Spanish, and German. Snowball uses a pre-defined set of stemming rules tailored to the specific morphological structure of each language, focusing on removing suffixes to reduce words to their root forms.

3.4 Test Collection Development

A reliable test collection is fundamental for evaluating the effectiveness of retrieval systems. The TREC-style approach, derived from the Cranfield paradigms, is widely used for developing test collections for LRLs [8, 37, 191].

Sahu and Pal [191] employed this method to create a test collection for Sanskrit using documents sourced from two online news portals. Query formulation and relevance judgment tasks were carried out by two Ph.D. students. The resulting test collection consists of 7,057 news articles and 50 topics, with an average of 8.54 relevant documents per query.

Similarly, Chavula and Suleman [37] developed a test collection for three Bantu languages (Chichewa, Citumbuka, and Cinyanja) using documents sourced from newspapers, Wikipedia, and web pages. Query development and relevance assessment were conducted by six recruited assessors. The collection includes 13,627 documents and 387 topics (129 topics for each language), distributed as: (1) 9,380 documents with an average of 19 relevant documents per query for Chichewa; (2) 2,258 documents with an average of 17 relevant documents per query for Citumbuka; and (3) 1,989 documents with an average of 15 relevant documents per query for Cinyanja.

Furthermore, AleAhmad et al. [8] created the Hamshahri test collection for Persian using a news corpus. This collection consists of 166,774 documents and 65 queries, with an average of 36.18 relevant documents per query. Query formulation and relevance assessment were performed by 17 university volunteer students.

3.5 Ad-hoc Text Retrieval

Researchers have investigated various retrieval and ranking strategies for ad-hoc text retrieval in LRLs, including the impact of language-specific features such as stopwords and stemming on retrieval effectiveness. Studies have reported mixed findings regarding the benefits of stopword

removal: some suggest that removing stopwords from both queries and documents enhances retrieval efficiency and effectiveness [191, 192], while others demonstrate that the impact of stopword removal depends on the ranking model and language, offering advantages in some cases but not in others [81, 100, 199].

For example, in a high-resource language like French, Savoy [199] found that retaining stopwords yielded better results than removing them when using the BM25 retrieval and ranking model. A similar trend was observed for the LRL Hindi, where retaining stopwords improved retrieval effectiveness, with a relative average gain of approximately 20% [81]. In contrast, no significant differences in retrieval effectiveness were reported for Marathi and Bengali.

Ghosh and Bhattacharya [100] reported variability in retrieval effectiveness across datasets within the same language. For instance, in Bangla, retaining stopwords achieved the highest MAP for the FIRE 2010 and 2011 datasets, whereas removing stopwords was more effective for the FIRE 2012 dataset. Similarly, for Hindi, retaining stopwords produced the best MAP for the FIRE 2010 dataset, while removing stopwords yielded better results for the FIRE 2011 dataset.

Regarding stemming, its effects on retrieval effectiveness also show variability across languages and dataset types. For French, Savoy [199] observed that applying stemming significantly improved retrieval effectiveness for short documents such as scientific abstracts (averaging 24.5 indexing terms), with improvements in average precision (AP) across various retrieval and ranking models, including BM25. However, for longer documents, such as news articles (averaging 182.2 terms per article), or when accents were ignored, stemming provided only marginal benefits. Similarly, Braschler and Ripplinger [29], using German data from CLEF 2000 and 2001 datasets, reported relative performance gains from stemming of up to 23% in MAP for short queries (title only) and up to 11% for long queries (a combination of title, description, and narrative).

The effects of stemming are inconsistent across languages. In a monolingual document retrieval experiment across eight languages using the CLEF 2002 dataset, Hollink et al. [115] found that stemming improved MAP for Finnish, French, German, and Swedish but had no positive impact on Dutch, English, Italian, or Spanish. Similarly, Flores and Moreira [93], using the CLEF 2005 and 2006 datasets, reported that stemming improved MAP for Portuguese, French, and Spanish but not for English when tested with different stemmers.

In studies focusing on Asian LRLs, Sahu and Pal [191] reported a 4.31% relative improvement in MAP for Sanskrit across six ranking models when stemming was applied. For Urdu, Sahu et al. [192] observed a 1.41% relative improvement using various stemming techniques. Similarly, Adriani et al. [2] reported a 2.00% relative improvement in MAP for Indonesian with stemming.

3.6 Summary

Web crawling is a widely used technique for building text corpora for LRLs, with a language identification model playing a key role in constructing corpora from web data. Other language processing tools, such as tokenizers, are equally essential for text preprocessing.

In ad-hoc text retrieval, preprocessing steps like stopwords removal and stemming are often applied to improve efficiency and effectiveness. However, studies show that their impact varies across languages, offering significant gains in some cases but limited or no benefit in others.

The effectiveness of retrieval systems is typically evaluated using test collections developed following TREC guidelines, which rely on human assessors for relevance judgments. Similar methodologies have been applied to LRLs to create test collections to enable the evaluation of retrieval systems.

To address the lack of essential resources for Tetun—such as a tokenizer, LID model, stopword list, stemmer, and test collection—we developed these tools and algorithms as part of this thesis. Their development is presented in Chapter 4, Chapter 5, and Chapter 6, respectively. These resources were then used to conduct experiments with various retrieval strategies, enabling the evaluation of ad-hoc text retrieval for Tetun, as detailed in Chapter 7.

Chapter 4

Tetun Text Resource Development

Language resources, such as text corpora, are essential for developing language technology components and tools, particularly in information retrieval (IR) and natural language processing (NLP). The lack of such resources for Tetun poses significant challenges to building technologies that effectively support the language. For LRLs, web crawling offers a practical solution by collecting documents from the web and then processing to build a text corpus [133, 146, 215, 235].

This chapter reports on two original papers published as part of the thesis research: the Labadain Crawler [74] and the Labadain-30k+ dataset [62], along with an analysis of the Labadain-30k+ dataset to examine how Tetun has evolved in digital spaces over time. The Labadain Crawler is a data collection pipeline designed to automate and optimize the construction of text corpora from the web for LRLs. As part of this pipeline, two Tetun-specific components were developed: a Tetun tokenizer [67] and a language identification model [66].

The Labadain Crawler pipeline [63] was successfully tested with Tetun [74] and used to build the Labadain-30k+ dataset [64], a monolingual Tetun text collection. To ensure quality, native speakers audited the dataset at the document level. Content analysis shows that it covers diverse sources and topics, making it suitable for a range of IR and NLP tasks.

4.1 Labadain Crawler Development

The Labadain Crawler [63] (*Labadain*, a Tetun word meaning “spider”) is a data collection pipeline designed to automate and streamline the construction of text corpora from the web for LRLs. It systematically manages the crawling, processing, and summarizing of textual data, enabling the creation of high-quality corpora. Built on Apache Nutch,¹ an open-source web crawling and data extraction framework, the pipeline provides a robust foundation for efficient corpus construction.

The automation of the Labadain Crawler relies on three essential components: an initial text corpus in the target language, a tokenizer, and a language identification model. These components are crucial for generating the seed words and URLs needed for Nutch to initiate the crawling process and for filtering text in the target language from the crawled data to construct a text corpus.

¹<https://nutch.apache.org>

4.1.1 Tetun Tokenizer

Given the absence of an existing tokenizer for Tetun, we developed a set of rule-based tokenization techniques specifically tailored to the language. Some of these techniques were designed for Tetun based on the standard of *Instituto Nacional de Linguística* (INL), while others were adapted from commonly used tokenization rules.

Due to the unique language-specific characteristics of Tetun, directly applying existing tokenizers is not feasible. For example, when using tokenizers for English, Portuguese, Spanish, and French from the NLTK library,² the Tetun sentence “*ha’u-nia uma mak ne’e*” (this is my home) is incorrectly tokenized. These tokenizers split the words “*ha’u*” and “*ne’e*” into [“*ha*”, “”, “*u*”] and [“*ne*”, “”, “*e*”], respectively. To address these challenges, a specialized tokenizer for Tetun was developed to ensure accurate tokenization.

4.1.1.1 Characteristics

The Tetun tokenizer [67] was developed using regular expressions and incorporates the following language-specific features:

- Tokens may include accented vowels and apostrophes, such as “*ne’ebé*” (which/that) and “*ne’ebá*” (there).
- Accented vowels can occur at the beginning, in the middle, or at the end of a token, for example, “*área*” (area), “*líder*” (leader), and “*oinsá*” (how).
- The accented consonant “*ñ*” typically appears within or in the middle of tokens, such as “*kompañia*” (company), and “*kampaíña*” (doorbell).
- Mono-semantic compound words may contain a combination of hyphens, accented vowels, and apostrophes, for instance “*sanulu-resin-ida*” (eleven), “*ida-ne’ebá*” (that one), and “*ida-ne’e*” (this one).
- Person names often include Portuguese-based accents, including the tilde, circumflex, acute, and cedilla, for example, “*João*”, “*Ângela*”, “*José*”, and “*Conceição*”

4.1.1.2 Techniques

The Tetun tokenizer applies several rule-based techniques, as outlined below:

1. *Word Tokenizer*: Extracts word units from the input text, excluding numbers, punctuation, and special characters.
2. *Simple Tokenizer*: Extracts words and numbers from the input text while discarding punctuation and special characters.

²<https://www.nltk.org>

3. *Standard Tokenizer*: Segments the input text into tokens based on word boundaries, punctuation, and special characters.
4. *Sentence Tokenizer*: Splits sentences using delimiters such as periods, question marks, and exclamation marks, while preserving periods used in abbreviations (e.g., Ph.D. and Dr.).
5. *Blank Line Tokenizer*: Segments the input text based on blank lines.

The **Word** and **Simple Tokenizer** are specifically designed to address Tetun’s unique linguistic features, while the other techniques are adaptations of commonly used tokenization approaches.

4.1.1.3 Evaluation and Result

The Tetun tokenizer was evaluated in a controlled experiment with five Timorese students, all native Tetun speakers. The goal was to assess its effectiveness on diverse Tetun text samples from the internet, including content from Wikipedia, reports, news articles, and other domains. To ensure consistency and reliability, the students followed a standardized guideline.

Each student was tasked with evaluating each tokenization technique by collecting at least three input texts from diverse internet domains, ensuring a range of linguistic features and structural complexities. The students then annotated the collected texts to create tokens that served as the gold standard for evaluating the tokenizer’s performance. An evaluation module was configured on their laptops, allowing them to input their sample texts and corresponding gold standard annotations to test the tokenizer. The resulting outputs were analyzed to assess the effectiveness of the techniques.

The evaluators reported that all Tetun tokenizer techniques achieved 100% accuracy for each tested input text, which consisted of 200 to 500 tokens sourced from various internet domains. Although the tokenizer was primarily developed based on the Tetun INL standard, its broader applicability to Tetun in general was also tested. This included evaluations using both Tetun INL-compliant texts and randomly collected Tetun texts from the web. These experiments demonstrated the tokenizer’s general applicability to Tetun and validated the effectiveness of its techniques, confirming its robustness and reliability for practical use.

4.1.2 Tetun Language Identification

Language identification (LID) is a fundamental task in NLP, enabling systems to classify and filter text based on a given language. For Tetun, the lack of an existing LID model limits the effort to build a text corpus. To address this gap, an LID model was developed to classify four languages commonly spoken in Timor-Leste: Tetun, Portuguese, English, and Indonesian.

4.1.2.1 Data Preparation

The text corpus for Tetun was sourced from Timor News [59], a platform dedicated exclusively to publishing news in Tetun, while datasets for Portuguese, English, and Indonesian were obtained

from the CC-100 dataset [235].³ Access to the Tetun text corpus was unrestricted, as it was obtained from a platform of the agency founded by the author of this thesis [56]. To ensure balanced representation, a subset of data equivalent in size to the Tetun text corpus was selected for Portuguese, English, and Indonesian from the CC-100 datasets using a random sampling technique.

The combined datasets were tokenized into sentences using the Tetun tokenizer, with each sentence labeled according to its respective language. Uniform preprocessing steps were applied across all datasets, including lowercasing and the removal of punctuation, special characters, and digits, to standardize the text for model training. A summary of the datasets used for training and evaluation is provided in Table 4.1.

Table 4.1: Summary of the dataset used for training and evaluating the LID model.

Description	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

4.1.2.2 Language Similarities Analysis

To analyze the similarities among Tetun, Portuguese, English, and Indonesian, the dataset was initially visualized using a Gaussian distribution plot. The interquartile range and standard deviation were then analyzed across parameter values ranging from 1 to 3. The analysis revealed that a standard deviation of 1 provided the most effective normalization for all input languages. This setting was adopted to filter the dataset to conform to the normal distribution range, which was subsequently used for clustering.

During clustering, input sentences were tokenized into word tokens using the Tetun Word Tokenizer, and the resulting tokens were converted into sparse Bag-of-Words vectors. These vectors were then subjected to dimensionality reduction via Principal Component Analysis (PCA), followed by feature normalization. The Gaussian Mixture algorithm was then applied to generate language clusters, selected for its effectiveness in LRL scenarios [83]. The resulting clusters were visualized in a two-dimensional space (see Figure 4.1), revealing clear dissimilarities among the four input languages.

4.1.2.3 LID Model Training

The LID model was trained using the dataset summarized in Table 4.1. Each sentence served as an input feature, while the target variable represented the corresponding language label. The dataset was divided into three subsets: 70% for training, 15% for development, and 15% for testing.

³<https://data.statmt.org/cc-100/>

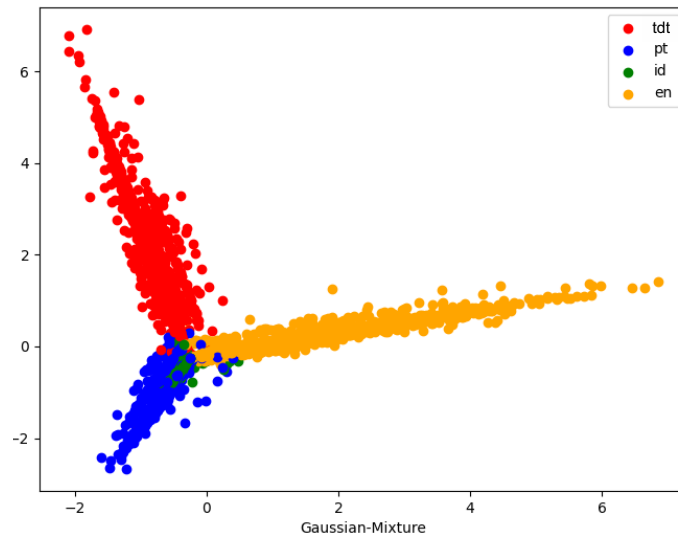


Figure 4.1: Gaussian-mixture clustering algorithm applied to cluster Tetun (tdt), Portuguese (pt), English (en), and Indonesian (id).

Experiments were conducted using three commonly employed machine learning models known for their strong performance in the LID task [85, 122]: support vector machine (SVM) with a linear kernel, logistic regression (LR), and multinomial Naive Bayes (MNB).

During training, the performance of these models was evaluated based on accuracy using the development set, with features extracted at both the character n-gram and word n-gram levels: (1) At the character n-gram level, each character sequence in the input text was treated as a feature, and (2) At the word n-gram level, each word served as a feature. The training results, including the comparative performance of the models across these feature levels when evaluated on the development set, are summarized in Table 4.2.

Table 4.2: Accuracy of the models' performance when evaluating using the development set.

Model	Character n-gram						Word n-gram		
	1	2	3	4	5	6	1	2	3
SVM	0.9822	0.9822	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

4.1.2.4 Model Evaluation and Results

The performance summary in Table 4.2 indicates that the MNB model with 5-gram character features achieved the highest accuracy (99.81%). Consequently, this configuration was selected for training and evaluating the LID model. When evaluating on the test set, the model achieved an overall accuracy of 99.77%. The F1 scores for individual languages were 99.87% for Tetun, 99.84% for Portuguese, 99.79% for Indonesian, and 99.76% for English.

4.1.3 Labadain Crawler Architecture

The architecture of the Labadain Crawler is presented in Figure 4.2. It consists of several key stages: providing an initial text corpus as input, generating seed words and URLs, crawling and indexing, text processing, and producing outputs (a final corpus and its summary).

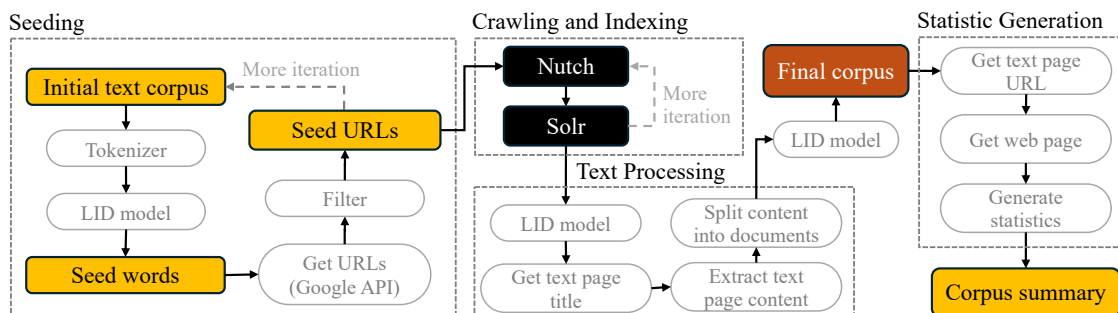


Figure 4.2: Architecture of the Labadain Crawler.

The modules for processing the initial corpus and generating seeds were adapted from techniques proposed by Baroni and Bernardini [20] and Linder et al. [146], with the addition of the LID model as an innovative layer introduced in this study. Crawling was conducted using Apache Nutch, and the outputs were automatically transferred and indexed in Solr.

The text processing and output generation modules represent novel approaches introduced to address the specific challenges of constructing a text corpus for LRLs. The following subsections present a detailed description of each functional stage of the Labadain Crawler.

4.1.3.1 An Initial Text Corpus

An initial text corpus is required to generate seed words. The corpus undergoes preprocessing, including lowercasing and removing punctuation, special characters, and digits. The preprocessed text is then tokenized into individual words, which are validated using the LID model to ensure that they meet the predefined threshold for the target language. The validated words form a vocabulary that is subsequently used in the seeding process.

4.1.3.2 Words and URLs Seeding

To generate a set of URLs, the vocabulary's frequency is calculated, and query seeds are created by randomly sampling three unique words, with sampling probabilities proportional to their frequency distribution. These words are combined to create a query, separated by spaces, and submitted to Google Search using the Google API.⁴

The top ten URLs for each query are retrieved and saved in Nutch's seed text file. To ensure quality, seed URLs are filtered to exclude duplicates and those linking to PDFs, Microsoft Office files, images, and audio or video files. This process of generating seed words and URLs is repeated based on the predefined configurations.

⁴<https://pypi.org/project/googlesearch-python/>

4.1.3.3 Crawling and Indexing

Using the provided set of URLs, Nutch sends HTTP requests to fetch the corresponding web pages and parses them to extract key elements such as links, metadata, and content. Nutch's link extractor identifies hyperlinks within the parsed pages, enabling the discovery of additional URLs for subsequent crawling.

The system maintains a queue of URLs to be crawled, iteratively fetching and parsing the pages within its frontier. A deduplication step at the content level removes duplicate documents. Once crawling is complete, Nutch automatically transfers the crawled web pages to Apache Solr for indexing, which organizes the content for efficient retrieval and analysis.

4.1.3.4 Text Processing

To retrieve text content in the target language, the Solr API is used to access the web pages indexed in Solr. The process consists of the following steps:

1. *Filter Web Pages:* Evaluate the title of each web page using the LID model. If the title's score meets or exceeds the predefined threshold, proceed to the next step.
2. *Extract Content:* Retrieve the content associated with the title, split it by newline characters, and store the resulting segments in a list.
3. *Filter Documents:* Apply the LID model to each segment in the list and filter out those that do not meet the predefined threshold based on their scores.
4. *Save the Results:* Save the segments that meet the threshold, together with their corresponding title and URL, into the final corpus file. Each extracted text page is separated by two consecutive newline characters.

4.1.3.5 Statistic Generation

To generate a summary for the final corpus, the following steps are performed:

1. *Load Corpus Content:* Read the content of the final corpus file.
2. *Extract URL Information:* For each content entry, extract the respective URL and count the occurrences of internet domain names and file types based on the URL extensions.
3. *Retrieve and Parse Web Pages:* Send HTTP requests to the URLs to retrieve the associated web pages. Use BeautifulSoup to parse the HTML content, enabling manipulation of the HTML elements.
4. *Analyze Links:* Extract inlinks (links to other pages within the same platform) and outlinks (links to external platforms) from each page, count the total number of links, and store this information in the corpus summary file. After the summary generation is completed,

additional statistics, such as the counts of domain names and file types, are also recorded in the corpus summary file.

4.1.4 Experiment on Constructing a Tetun Text Corpus

The Labadain Crawler was tested with Tetun using the following experimental settings:

1. *Input Text Corpus:* A corpus was created by randomly selecting 10% of the 3,500 Tetun news articles extracted from Timor News.
2. *Tokenization and Filtering:* The Tetun Word Tokenizer was applied to segment the input text into word tokens. The Tetun LID model, with a score threshold of 0.95, filtered the candidate words.
3. *Seeding and Crawling:* The seeding process was configured to repeat ten times, and Apache Nutch (version 1.19, released in September 2022) was set to use a depth-first crawling strategy with a depth limit of five.
4. *Crawling and Indexing:* The crawling process was configured to repeat fifteen times, with automated indexing in Solr (version 8.9.0, released in June 2021, the latest available version at the time of the experiment). Solr was configured according to the official Nutch documentation,⁵ using the default settings.

Once the crawling process was initiated, all subsequent steps were executed automatically until the construction of the text corpus and the generation of its summary were completed. The experiment was conducted on a Linux Ubuntu virtual machine equipped with 16GB of RAM, 150GB of hard disk space, and a single socket with four cores. The entire process was completed in approximately 46 hours.

4.1.5 Experimental Result Analysis

In the seeding stage, ten iterations were performed to generate seed words and URLs, resulting in a total of 41 domains and 61 unique URLs. These URLs were used as input seeds for Nutch to initiate the crawling process. Upon completion of the crawling step, text processing was performed, resulting in the creation of the final text corpus and its corresponding summary. The top five internet domains from which Tetun text pages were retrieved are presented in Table 4.3, and the overall corpus summary is provided in Table 4.4.

The corpus obtained from the crawled data was sourced from 149 distinct internet domains, with each text page representing the plain text content of a web page. The content of each crawled web page was parsed and structured within Nutch, producing a document suitable for indexing in Solr. Each document includes metadata fields such as title, URL, and content. Some text pages contain multiple articles, each linked to its respective title and URL. The corpus also includes text

⁵<https://cwiki.apache.org/confluence/display/NUTCH/NutchTutorial>

Table 4.3: Top five domains in the dataset.

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

Table 4.4: Summary of the corpus content (excluding titles and URLs). *Tokens do not include punctuation and special characters.

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

extracted from 589 PDF files and 13 PowerPoint files, with an average of 142.12 inlinks and 14.13 outlinks per URL.

Based on the internet domain names, a comprehensive analysis was conducted, and the data sources were manually categorized into the following groups:

1. *Online Newspapers*: News portals affiliated with news agencies.
2. *Governmental Institutions*: Online portals maintained by government bodies.
3. *Non-Governmental Institutions*: Platforms associated with national and international non-governmental organizations and agencies.
4. *Educational Institutions*: Online portals linked to universities and educational institutions.
5. *Blogs and Forums*: Platforms for personal opinions and discussion forums.
6. *Personal Pages*: Portals showcasing individuals' personal information and interests.
7. *Wikipedia*: Documents sourced from the Tetun Wikipedia.
8. *Banks and Courts*: Platforms associated with banks and judicial institutions.

The summary of the corpus by data source category is presented in [Table 4.5](#), and the distribution of text pages by domain name is summarized in [Table 4.6](#).

4.1.6 Evaluation of the Corpus Quality

To ensure that the text pages collected by the Labadain Crawler were Tetun documents containing high-quality data, we evaluated the corpus using both the summary and final corpus files. The

Table 4.5: Total number of text pages per data source.

Data source category	#text pages	Proportion
Online newspapers	16,509	73.73%
Government institutions	3,222	14.39%
Non-government 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 4.6: Total number of text pages per top-level domain.

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

initial analysis focused on the summary file, which lists domain names along with the number of text pages per domain. This provided insights into the overall composition of the corpus and helped identify non-Tetun texts.

The analysis revealed that among the 149 domains listed in the corpus summary file, 22 domains were flagged as potentially containing non-Tetun text pages. To confirm this, we manually cross-referenced the corpus content with the corresponding domain names. The analysis revealed that 20 of these 22 domains did not contain Tetun text pages. These 20 domains accounted for 22 text pages, representing approximately 0.01% of the corpus content. Since the LID model was applied during the construction of the final corpus, these non-Tetun pages resulted in empty content (i.e., they lacked body text) and were subsequently excluded from the final corpus.

To further evaluate the quality of titles and content in the final corpus file, we adopted the “quality at a glance” technique recommended by Kreutzer et al. [134] and employed by Artetxe et al. [16]. A random sample of 50 text pages was selected from the corpus and distributed to six native Tetun-speaking students, resulting in a total of 300 text pages being assessed. Evaluators were provided with access to the sample corpus and Google Forms to facilitate the evaluation process. They assessed several aspects of corpus quality, including the quality of text page titles, noise levels, recency and relevance of the content, and the overall quality of the corpus. The detailed assessment criteria are provided in Section A.2 of Appendix A.

The evaluation results showed that out of 300 text pages, five lacked body text, as shown in Table 4.7. An in-depth analysis determined that these pages contained only layout elements such as menu names or links, without actual body content. This issue arose because these pages were linked to scanned PDF documents, web home pages, or pages without Tetun body text, despite

having titles in Tetun. Overall, the qualitative evaluation demonstrated that the generated corpus consisted of high-quality textual data from diverse sources, with clean titles and content written in Tetun. These findings demonstrate the effectiveness of the Labadain Crawler in collecting Tetun documents from the web.

Table 4.7: Results of the corpus quality assessment.

Quality metric	Description	#text pages	Proportion
Text page title quality	The text page title is in Tetun	300	100%
Text page content quality	The text page contains one or more articles	295	98.33%
Noise	The text page contains clean text	300	100%
Recency and relevancy	Relevant content for present-day usage	278	92.67%
Overall assessment	Diverse sources with high-quality content	295	98.33%

4.1.7 Discussion

The Tetun Word Tokenizer, although specifically designed to address tokenization challenges in Tetun, has demonstrated adaptability by effectively processing texts in other languages, including English, Portuguese, and Indonesian. This capability became evident during the analysis of language similarities, where the tokenizer successfully tokenized texts from these languages into word tokens. Its versatility significantly enhances its utility, making it a valuable resource for a broader range of linguistic contexts beyond its original focus on Tetun.

Although previous studies on Tetun emphasize the influence of Portuguese loanwords, suggesting they account for up to 40% of the language [103, 106, 223], the clustering results presented in Figure 4.1 do not indicate notable similarities between Tetun and Portuguese. This finding was further supported by the analysis and assessment of corpus quality. A random inspection of text page titles in the corpus revealed that the LID model accurately classified some titles dominated by Portuguese loanwords as Tetun. Examples include “*CCLN Selebra Akordu Apoiu Grupu Veteranu*” and “*MTK Relata Progresu Investimentu Infrastrutura*”, where all words, except “CCLN” and “MTK”, are not Portuguese loanwords.

However, the LID model exhibited biases toward specific terms, such as “COVID” and “Timor-Leste”. These biases likely derive from the higher frequency of these terms in the Tetun text corpus used for training the LID model, while their presence in the datasets of the other three languages was limited or absent. Consequently, when these terms appeared in short texts or as standalone instances, the model tended to classify them as Tetun. This bias was evident in the 20 domains associated with 22 text pages identified during the corpus analysis by internet domain names. These pages, which typically had short titles consisting of one to four words, included either “COVID”, “Timor-Leste”, or both, leading to their incorrect classification as Tetun.

Despite these challenges, the Labadain Crawler demonstrated its effectiveness in collecting Tetun documents from the WWW, even with limited computational resources. It successfully operated on a personal computer with a minimum configuration of 16GB of RAM and 150GB of

hard disk space. Additionally, the pipeline generated a summary of the crawled data, providing valuable insights into the corpus’s characteristics, the distribution of documents across internet domains, and the diversity of the information collected.

In a multilingual environment like Timor-Leste, challenges commonly arise from mixed-language texts, which are often present within the same document. In some cases, document titles may be in one language while the content is in another, leading to empty document bodies after applying the LID model. This highlights the importance of auditing document content to ensure the quality of the web-crawled data.

4.1.8 Conclusions

The Labadain Crawler, built on Apache Nutch, was successfully integrated with a tokenizer and a language identification (LID) model developed specifically for Tetun. These components, created as part of this research, enable effective language-specific processing. Additionally, custom modules for text processing and statistics generation were implemented, representing original contributions introduced in the Labadain Crawler publication [74].

When tested with Tetun, the crawler successfully constructed a text corpus comprising 321,721 sentences extracted from 22,392 web pages. Each page represents plain text retrieved from the web, underscoring the crawler’s effectiveness in acquiring and processing Tetun documents.

A manual assessment of the corpus confirmed both the diversity of sources and the high quality of the content, demonstrating its suitability for a wide range of NLP and IR tasks. These results validate the crawler’s capability to collect high-quality Tetun data from the web. Furthermore, its modular design ensures adaptability and ease of extension, making it applicable to other low-resource languages facing similar challenges in corpus construction.

4.2 Labadain-30k+ Dataset Construction

Advancing language technology tools requires high-quality datasets. To build a Tetun text corpus, we used the Labadain Crawler [63] to collect a substantial volume of web data as the foundation for a Tetun dataset. This dataset was further enriched with Tetun documents from the multilingual Wikipedia dump released by the Wikimedia Foundation [171]. By combining these sources, processing the data, and conducting quality audits with native Tetun speakers, we developed the Labadain-30k+ dataset [64], a resource designed to support Tetun NLP and IR tasks.

The web-crawled data includes titles, URLs, and plain text, encompassing a variety of content elements such as headings, subheadings, links, body content, and user comments. The Wikipedia documents provide structured information, including IDs, titles, URLs, and content, complementing the web-crawled data by contributing additional document diversity and structured text.

The following subsections present a detailed description of the process of constructing the Labadain-30k+ dataset. This includes the document annotation process, content extraction and processing, document characterization, and the incorporation of Wikipedia documents from Wikimedia to create a text dataset for Tetun.

4.2.1 Document Annotation

To enhance the dataset’s suitability for a broad range of IR and NLP tasks, a thorough document-level annotation process was conducted to ensure data quality. The web-crawled documents were manually audited by two volunteer linguists—recent graduates specializing in the Tetun language. Their primary task was to analyze the content and identify key structural elements, such as body sections and publication dates, for each domain, using the associated URLs as references.

4.2.1.1 Annotation Processes

The annotation process is illustrated in Figure 4.3. Initially, internet domain names were automatically extracted from the URLs in the raw text data. Documents were then grouped by domain, with each group saved into a separate file. This step produced a total of 79 domains, 26 of which contained more than 100 documents each.

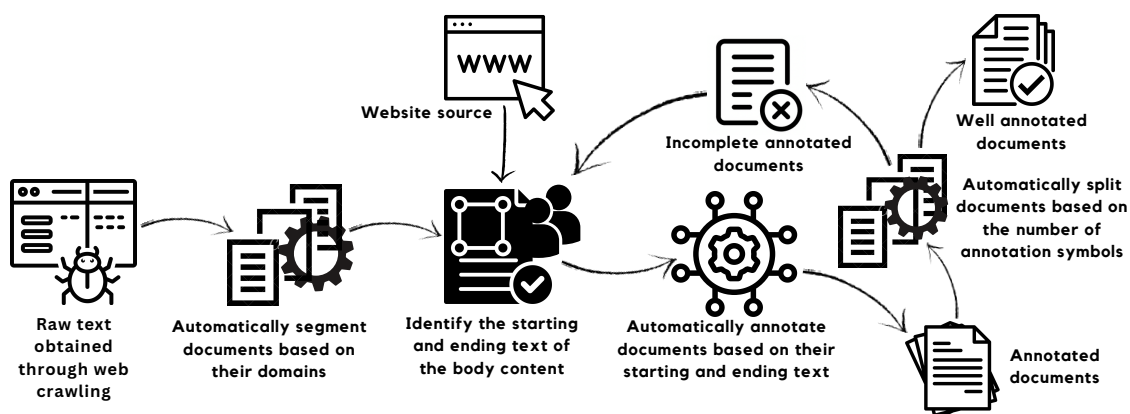


Figure 4.3: Document annotation process flow.

Before annotating the documents, annotators were tasked with analyzing the structure of web pages and the format of publication dates for each domain by referencing the source website accessed through the URL provided in the documents. Based on this analysis, the annotators identified a set of potential `start` and `end` text patterns for each domain, corresponding to its page layout. These patterns were then used to automate the content annotation process, as described in the algorithm detailed in Section A.3 of Appendix A.

The automated annotation process produced an output saved in the `annotated documents` file, as shown in Figure 4.3. An automated verification step was subsequently performed to assess the quality of the annotated documents. Documents containing exactly two `<t>`⁶ tags were classified as well-annotated and recorded in the `well-annotated documents` file. Documents that did not meet this criterion were saved in the `incomplete annotated documents` file for further review.

⁶The `<t>` tag was chosen as a notation preference by the authors and can be replaced with any other annotation tag as needed.

The incomplete annotated documents were returned to the annotators for further analysis to address annotation issues. Annotators revised the `start` and `end` text patterns as needed, reapplied the content annotation algorithm, and repeated the process iteratively until the documents were correctly annotated. For documents extracted from PDFs and presentation files, manual annotation was performed before adding them to the `well-annotated documents` file.

4.2.1.2 Publication Dates Identification

The identification of publication dates followed a two-step approach. The first step involved analyzing each URL to determine whether it embedded a publication date. The second step required manually browsing the corresponding web pages to check whether publication dates were present within the content. For documents extracted from PDF or presentation files, dates found within the files themselves were used. In cases where a single page contained multiple documents with different dates, the publication date displayed at the top of the page was applied to all documents on that page.

Annotators provided detailed instructions for locating publication dates specific to each domain. If the URLs did not include publication dates, additional guidance was given on identifying date formats within the page content. This information was incorporated into the configuration used to extract publication dates from documents for each domain. All extracted publication dates were standardized according to the ISO 8601 format [117].

4.2.2 Content and Date Extractions

The content extraction process was automated using the `well-annotated documents` file as input. Only the text enclosed between `<t>` tags was extracted, while all other content was discarded. For publication date extraction, a heuristic was applied to automate the process when the document's URL contained an embedded date.

For documents where URLs did not include publication dates, the corresponding website was manually inspected to identify the CSS class tags associated with the publication date. After compiling the relevant tags and date formats for all domains, BeautifulSoup was employed to automatically extract publication dates for each document. The extracted publication dates, along with the title, URL, document source, and content, were then saved in the output file.

An additional automated verification process was performed to ensure uniformity in date formats and structure. Any incomplete or inconsistent information was iteratively corrected to maintain a consistent structure across all documents.

4.2.3 Deduplication and Post Processing

The deduplication process involved validating both document titles and their corresponding URLs to identify and exclude duplicate entries. Repeated instances of document titles embedded within the content were eliminated to improve consistency. To further enhance document quality, occurrences of document source names within titles were identified and deleted using a find-and-replace

function. For example, “| *Notísia Timor News*” was eliminated from the title [Povu mak sei Hili | **Notísia Timor News**].

The finalized dataset comprises 32,113 documents, each containing a title, URL, document source, publication date, and content. A detailed dataset summary is presented in [Table 4.8](#).

Table 4.8: Summary of the web-crawled dataset.

Data source	#documents	Proportion
Online newspapers	28,997	90.30%
Non-government 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%

4.2.4 Document Characterization

To ensure diversity among the documents, a subset of 2,879 documents—sourced from non-governmental, governmental, educational, banking, and court portals—was selected for the characterization task. These documents, highlighted rows of [Table 4.8](#), were chosen for their diverse content representation. Through a preliminary analysis, the following seven categories were identified and incorporated into the document characterization guidelines:

1. *News Articles*: Informational content, including updates on activities, events, and remarkable occurrences.
2. *Legal/Governmental Documents*: Materials related to judiciary matters and governmental affairs, comprising legal texts, legislation, official communications, and government-related publications. Examples include Council of Ministers meeting notes, press releases, speeches by government officials, profiles of government members, and parliamentary documents.
3. *Technical Documents*: Documents containing technical information, instructions, and specifications across various fields such as science, engineering, technology, politics, and finance. Examples include reports, presentations, and policy recommendations.
4. *Correspondence Letters*: This category comprises personal and professional communications, including letters, emails, and other written exchanges between individuals or entities.
5. *Research Papers*: Documents presenting research findings, analyses, and insights on specific topics. Examples include research articles, conference papers, position papers, review papers, survey papers, case studies, and related materials.

6. *Institutional Information*: Documents providing details about institutions, such as universities, colleges, banks, government agencies, and non-governmental organizations.
7. *Advertisements/Announcements*: Documents intended to promote products, services, events, tenders, job vacancies, or to make public announcements. Examples include advertisements, promotional materials, and similar information.

4.2.4.1 Characterization Processes

The document characterization task was carried out by three native Tetun-speaking students following established guidelines. The document characterization process began with annotators reviewing the guidelines to understand the task requirements. Annotators were instructed to familiarize themselves with the predefined categories by analyzing examples of documents in each category provided in the guidelines.

Subsequently, a training session was provided to demonstrate practical annotation examples. Annotators then participated in three pilot testing sessions, each involving the characterization of ten documents. After completing each session, annotators compared their results, discussed challenges encountered, suggested improvements, and incorporated feedback to enhance the accuracy of the document characterization.

After completing the training and pilot sessions, each annotator independently characterized the 2,879 documents. The task was completed within two days, totaling approximately 16 hours, with an average characterization time of 20 seconds per document.

4.2.4.2 Inter-Annotators Agreement

To evaluate the reliability of inter-annotator agreement, Fleiss’s Kappa measure [92] was used, with the strength of the agreement interpreted based on the table of Landis and Koch [139]. The analysis resulted in a k value of 0.4994, indicating moderate agreement among the annotators.

Following the analysis, the annotators reviewed their discrepancies, discussed differences, and reached a consensus for all documents. The finalized document distribution comprises 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.

4.2.5 Wikipedia Document Processing

To augment the web-crawled data, we incorporated Tetun Wikipedia documents from Wikimedia [171], available on Hugging Face.⁷ The extraction process followed the dataset’s official guidelines and produced 1,468 documents, each containing an ID, URL, title, and content.

To ensure consistency with the previously web-crawled data, the same methods outlined in Subsection 4.2.2 were applied. This included generating document sources, extracting publication

⁷<https://huggingface.co/datasets/wikimedia/wikipedia>

dates, and organizing the content to match the format of the crawled data, with each document separated by two consecutive newlines. The preprocessing steps involved removing HTML tags present in certain documents and excluding document IDs from the dataset.

The documents were then distributed to the three previously mentioned students for content auditing, with each reviewing approximately 500 documents. During the review process, 13 documents were identified as either empty or non-Tetun documents. Additional issues, such as mixing Tetun with Indonesian and English, were also observed. These problematic documents were eliminated during auditing, resulting in a final dataset of 1,455 clean documents.

4.2.6 Final Dataset

The final dataset was compiled by combining 29,234 uncharacterized documents (corresponding to the non-highlighted rows in Table 4.8) with the 2,897 consensus documents described in Subsection 4.2.4 and the 1,455 Wikipedia documents detailed in Subsection 4.2.5.

At this stage, 18 duplicate documents were identified in the Wikipedia set. After deduplication, 1,437 unique documents were retained and combined with the 32,113 documents described in Subsection 4.2.3, resulting in a final dataset of 33,550 documents, referred to as the **Labadain-30k+** dataset [64]. Each document in the dataset comprises metadata such as a title, URL, document source, document category, publication date, and content. A detailed summary of the final dataset is provided in Table 4.9.

Table 4.9: Summary of the final dataset.

Category	#documents	Proportion
News articles	30,150	89.87%
Wikipedia documents	1,455	4.34%
Legal/government documents	1,223	3.65%
Technical documents	211	0.63%
Blogs and Forums	145	0.43%
Advertisements/announcements	124	0.37%
Research papers	83	0.25%
Personal pages	74	0.22%
Institutional information	53	0.16%
Correspondence letters	32	0.10%

The Labadain-30k+ dataset is publicly available in the INESC TEC repository under a Creative Commons Attribution-ShareAlike (CC BY-SA) license. It is provided as a single text file (84 MB in size), with documents separated by two consecutive newlines.

4.2.7 Discussion

The total number of documents in the Labadain-30k+ dataset is comparable to the number of Tetun documents in MADLAD-400 [136], a multilingual dataset released by the Google Research and Google DeepMind teams in October 2023, with Labadain-30k+ containing approximately 6.8k

fewer documents. While both datasets compile Tetun documents from the web, their collection methods and document structures differ significantly.

The Labadain-30k+ dataset was created using the Labadain Crawler [63], developed as part of this thesis, and includes documents up to September 30, 2023. It provides detailed metadata such as URLs, document sources, publication dates, categories, and content, making it particularly valuable for tasks such as information retrieval, document classification, and metadata-driven analyses. In contrast, MADLAD-400 was constructed from Common Crawl snapshots collected between 2008 and August 2022. The Tetun documents in MADLAD-400 are available only as plain text, without the metadata information provided in the Labadain-30k+ dataset.

Additionally, the Labadain-30k+ dataset was audited by native Tetun speakers, ensuring higher data quality, accuracy, and linguistic relevance. In contrast, MADLAD-400 did not undergo similar auditing, which may result in less accurate Tetun text. This distinction highlights the Labadain-30k+ dataset’s potential for applications that require high-quality text data.

Furthermore, Table 4.10 compares the size of the Labadain-30k+ dataset with the number of speakers for other LRLs, with total document counts based on the statistics reported in MADLAD-400. Tetun, Occitan, and Mizo exhibit similar dataset sizes available on the web and a comparable number of speakers. Notably, despite Tetun having a smaller speaker base (932k+) compared to Assamese (15M+) and Swiss German (5M+), its dataset size (33.6k documents) is nearly equivalent to Assamese (33.8k documents) and relatively close to Swiss German (42.7k documents).

Table 4.10: Comparison of the Labadain-30k+ dataset size and total number of speakers with other low-resource languages.

Language	#documents	#speakers
Tetun	33.6k	932k+
Assamese	33.8k [136]	15M+ [30]
Occitan	36.4k [136]	1.5M [179]
Mizo	36.4k [136]	~1M [220]
Swiss German	42.7k [136]	5M+ [214]

This comparison highlights the potential of Tetun as a well-represented LRL in terms of available data despite its relatively smaller speaker population. It demonstrates that even languages with limited speaker bases can achieve substantial representation, positioning datasets like Labadain-30k+ as valuable resources for advancing Tetun IR and NLP tasks research.

4.2.8 Conclusions

The Labadain-30k+ dataset marks a crucial milestone in Tetun-language resources, serving as the first-ever Tetun text dataset audited by native speakers. Consisting of 33,550 documents, the dataset is enriched with metadata such as URLs, document sources, publication dates, categories, and content. Its diverse composition, covering a wide range of sources and categories—such as

news articles, legal texts, research papers, and technical documents—enhances its applicability for various research domains.

The inclusion of rich metadata provides valuable contextual information for each document, making the dataset highly suitable for IR, search-related research, and various NLP tasks. Its structured format and document diversity enable researchers to address challenges associated with Tetun while exploring innovative approaches to improving search solutions for the language.

This study introduces scalable methodologies for document annotation and characterization, designed to ensure the quality and usability of the resulting dataset. These methodologies offer a replicable framework for constructing high-quality textual corpora from the web for other LRLs. By addressing challenges such as limited computational resources, this work contributes to the broader efforts of advancing language technologies for underrepresented languages.

4.3 Labadain-30k+ Dataset Analysis

To explore the evolution of Tetun text on the web, this section analyzes document composition, publication trends over time, the adoption of the official Tetun INL linguistic standard, and the influence of Portuguese loanwords. The analysis is based on the Labadain-30k+ dataset and aims to assess how Tetun has developed in digital spaces over time.

To ensure consistency, the terms used are defined as follows: (1) *Document*: A dataset unit comprising a title, URL, source, publication date, and content. (2) *Title*: Title of the document. (3) *Content*: The body text of the document. (4) *Corpus*: A combination of document titles and their content. (5) *Paragraph*: A segment of text separated by a single newline within the document content. (6) *Sentence*: A line of text ending with a period, exclamation mark, or question mark. Periods in abbreviations, such as Dr. or Ph.D., are not treated as sentence delimiters. (7) *Token*: A unit of text consisting of a word or number, excluding punctuation and special characters. (8) *Vocabulary*: A set of unique tokens.

4.3.1 Dataset Description and Distribution

A quantitative description of the Labadain-30k+ dataset is presented in Table 4.11, while Table 4.12 provides detailed statistics, including the number of paragraphs, sentences, and tokens.

Table 4.11: Description of the Labadain-30k+ Dataset.

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 4.12: Summary of the documents in the Labadain-30k+ dataset.

Description	Minimum	Maximum	Average
Total number of paragraphs	1	1,109	9.98
Total number of sentences	1	936	12.35
Total number of tokens (titles)	1	29	9.15
Total number of tokens (content)	2	27,166	357.48

To identify the primary contributors to the Labadain-30k+ dataset and their respective origins, the documents were grouped by source. The analysis reveals that the top five contributors, based on document count, predominantly originate from online newspapers, as shown in Table 4.13.

Table 4.13: Top five sources by document count.

Source	#documents	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%

Among these, Tatoli, the public online news agency in Timor-Leste, is the largest contributor, accounting for 27.19% of the dataset, followed by Timor Post with 13.79%. The remaining sources in the top five are Naunil, Tempo Timor, and Timor News.

To understand the dataset’s composition, documents were grouped by top-level domains (TLDs), as presented in Table 4.14. The “.com” domain was the most dominant, followed by “.tl”, the country code TLD for Timor-Leste.

Table 4.14: Summary of the dataset per TLDs.

Source	#documents	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%

The word frequency distribution within the corpus was examined to assess its adherence to Zipf’s law [256]. The plot in Figure 4.4 illustrates the relationship between word rank and frequency, confirming the expected Zipfian distribution. This relationship reflects the inverse proportionality between a word’s frequency and its rank. Excluding stopwords, the most frequent words in the corpus include terms such as “*governu*” (government), “*timor-leste*”, and “*dili*” (capital of Timor-Leste), highlighting key terms prevalent in the dataset.

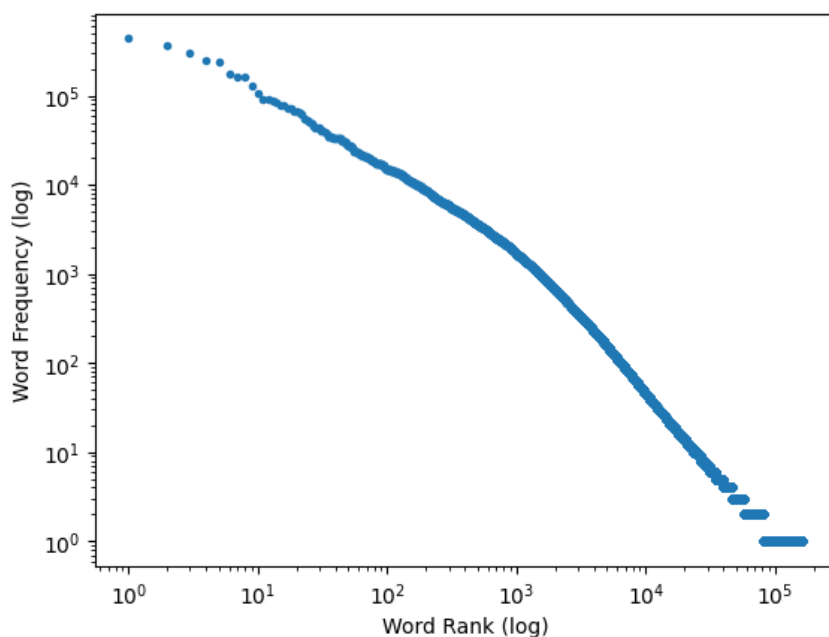


Figure 4.4: Word frequency vs. Word rank.

Furthermore, an analysis of co-occurring word sequences was conducted, focusing on bigram and trigram words as representative examples. The bigram words analysis revealed prominent word pairs such as “*covid 19*” and “*prezidente república*” (president of the republic). Similarly, the trigram analysis identified frequent patterns like “*taur matan ruak*” (name of the prime minister of Timor-Leste from 2018 to 2023) and “*guterres lú olo*” (name of the president of Timor-Leste from 2018 to 2023). These n-grams underscore recurring topics and entities within the dataset.

4.3.2 Document Evolution on the Web

Analysis of the document evolution reveals that the Labadain-30k+ dataset comprises documents spanning from 2001 to 2023, with no documents available for 2004 and 2005. The absence of documents from these years may be attributed to factors such as language barriers and limited digital archiving efforts due to constraints in internet infrastructure [118, 164]. The dataset contains fewer than 100 documents for years before 2010, reflecting similar challenges during that period.

Starting in 2017, there was a substantial increase in the number of documents, as presented in Table 4.15, driven by the growing popularity of online news. This growth can be associated with the launch of Tatoli in March 2017 [170]. A sustained rise in the number of web documents became noticeable only from 2020 onward, continuing consistently through 2023.

To assess the evolution of document writing quality, the analysis focused on two key aspects: adherence to the Tetun INL standard and the influence of Portuguese loanwords. Each document was tokenized into word tokens, and each token was cross-referenced with the INL dictionary [49] to evaluate adherence to official Tetun writing norms. In parallel, tokens were also compared against Greksakova’s dictionary [103] to identify and measure the presence of loanwords.

Table 4.15: Yearly evolution of document quantity. ⁺Percentage points.

Year	#documents	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

Each word was first checked against the INL dictionary to determine whether it conformed to the official Tetun standard. It was then cross-checked with Greksakova's dictionary to identify Portuguese loanwords. Words not found in either dictionary were classified as misspelled or out-of-vocabulary (OOV). Additionally, given the substantial increase in web document quantity since 2017, this year was selected as a threshold to compare the evolution of Tetun writing quality and loanword usage before and after 2017.

The analysis revealed a significant improvement in adherence to the Tetun INL standard, alongside a continued prevalence of Portuguese loanwords (see details in Table 4.16). Additionally, the proportion of misspelled and OOV words declined notably after 2017, further highlighting improvements in writing quality.

Table 4.16: Evolution of Tetun INL adoption in document writing before and after 2017. ⁺Numbers are excluded from the count. ^{*}Some loanwords are also included in the Tetun INL dictionary.

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

Given the predominance of news articles, a focused analysis was conducted on the evolution of document writing content and the impact of Portuguese loanwords within this category. The findings revealed that news articles exhibit a higher adherence to the Tetun INL standard and a higher frequency of Portuguese loanwords compared to the overall dataset (see Figure 4.5).

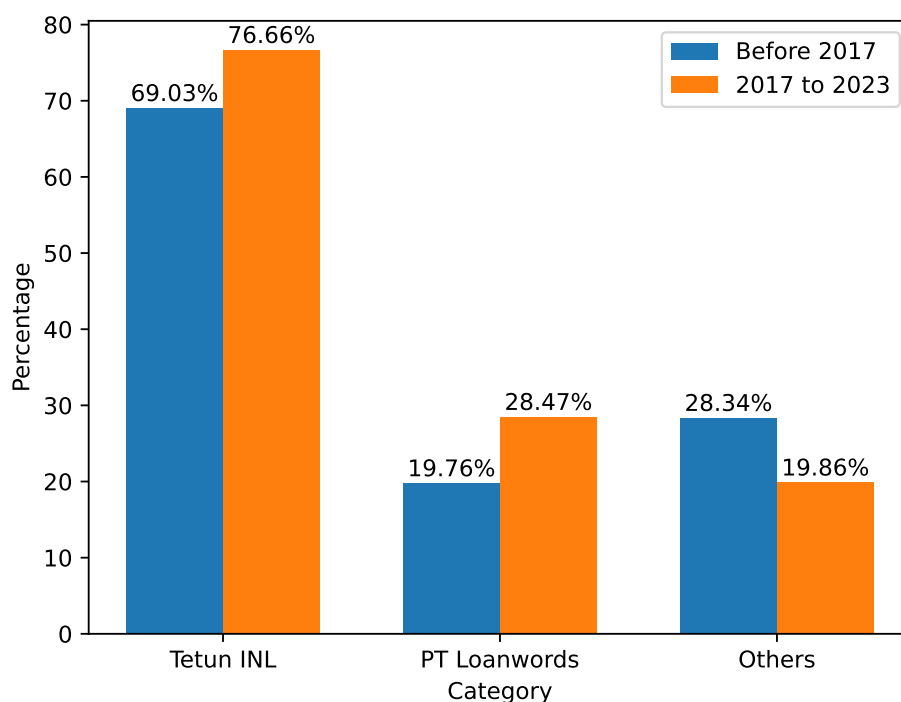


Figure 4.5: Evolution of document writing adherence to the Tetun INL standard and the influence of Portuguese loanwords in news articles before and after 2017.

4.3.3 Analysis on the Results

The Labadain-30k+ dataset includes documents from diverse sources (Table 4.8) and multiple categories (Table 4.9). While news articles constitute the majority (Table 4.13), Wikipedia and legal/government documents also contribute substantially. Smaller contributions are spread across seven additional categories, each with fewer than 300 documents. An analysis of document origins by TLD shows that most documents are hosted on “.com” domains, followed closely by “.tl”, domains, with a margin of just 2.56 percentage points (Table 4.14).

From a linguistic perspective, word frequency distribution in the Labadain-30k+ dataset follows Zipf’s law, emphasizing the concept that a small number of words occur frequently, while the majority exhibit lower frequencies (Figure 4.4). Additionally, an analysis of up to 5-gram words, excluding stopwords, reveals that a substantial portion of the documents focus on the COVID-19 pandemic, events in Dili, and topics related to the country and government activities.

In terms of web-based document growth, a steady increase has been observed since 2012, with a notable surge occurring in 2020—an 11.75 percentage point rise compared to 2019 (Table 4.15). This upward trend continued through 2023. However, due to the dataset cutoff at September 30, 2023, a 3.79 percentage point decrease was observed for that year compared to 2022.

Regarding writing quality, evaluated against the Tetun INL standard and focused on vocabulary usage, the analysis shows a +6.12 percentage point increase in standard adherence from 2017 onward. Meanwhile, the use of Portuguese loanwords remained significant, with a +5.09 percentage point increase over the same period (Table 4.16).

4.3.4 Discussion

The significant increase in document quantity from 2020 onward, combined with the prominence of “covid 19” as the most frequent word pair, suggests a strong correlation between the COVID-19 pandemic and the rise of Tetun documents on the web. Approximately 90% of these documents are news articles, which exhibit a notable improvement in writing quality—specifically greater adherence to the Tetun INL standard—since 2017. This improvement surpasses the overall dataset trend by +1.51 percentage points (Figure 4.5). Furthermore, the use of Portuguese loanwords in news articles exceeds the overall improvement by +3.62 percentage points, underscoring the influential role of online news in shaping both the adoption of the Tetun INL standard and the integration of Portuguese vocabulary into written Tetun documents.

Previous studies reported a five percentage point increase in the prevalence of Portuguese loanwords in Tetun newspapers, rising from 35% to 40% between 2018 and 2019 [103, 106], with some news titles predominantly composed of Portuguese loanwords. To further investigate this trend, a comparative analysis was conducted using news article titles from the same periods. The results revealed a similar upward trend, though with a marginal increase of +3.5 percentage points and a lower overall proportion of Portuguese loanwords: 30.01% in 2018 and 33.51% in 2019. While differences may result from dataset variations, the observed increase aligns with prior findings, reinforcing the trend of growing Portuguese influence in Tetun news writing.

As shown in Table 4.16, 20.23% of words in the dataset were not found in dictionaries and were classified as misspelled, out-of-vocabulary (OOV), or from other languages. These often include domain-specific terms or named entities. An analysis of the top 10 most frequent words in this category identified terms such as “*hanesan*” (such as), “*Timor-Leste*”, “*hetan*” (get), *PNTL* (National Police of Timor-Leste), and “*Covid*” as OOV words. The absence of such frequent terms from the Tetun INL dictionary highlights a key limitation, underscoring the need to expand and update lexical resources to better support language processing tasks.

4.3.5 Conclusions

The analysis of the Labadain-30k+ dataset provides valuable insights into the evolution of Tetun documents on the web, with approximately 90% of the documents originating from online news. The increase in document quantity from 2020 onward reflects the growing digital presence of Tetun documents, significantly influenced by the COVID-19 pandemic. News articles show notable progress in adhering to the Tetun INL standard, while continuing to incorporate a consistent level of Portuguese loanwords.

Despite these advancements, the findings also reveal ongoing challenges, particularly the prevalence of OOV words. Many commonly used terms are not represented in the INL dictionary, revealing a critical gap in available language resources. Expanding and updating these resources is crucial to better support contemporary Tetun usage and accommodate evolving linguistic nature.

This study contributes a comprehensive analysis of Tetun’s digital presence, identifies key gaps in linguistic resources, and highlights the need for further standardization and vocabulary

expansion. The findings underscore the importance of updating and extending the Tetun INL dictionary to more accurately reflect current usage, particularly in digital contexts where OOV words can limit language processing effectiveness.

More broadly, the methodology presented in this work offers a replicable framework for other LRLs. It demonstrates how systematic resource development and dataset analysis can help address linguistic gaps, promote language standardization, and advance both NLP and IR for languages with limited digital representation.

Chapter 5

Tetun Language Component Building

Language components, such as stopword lists and stemmers, are essential for supporting text-processing tasks in information retrieval (IR) and natural language processing (NLP). In IR, these components typically enhance both retrieval efficiency and effectiveness. The absence of such resources for Tetun has hindered advancements in Tetun text IR.

For low-resource languages (LRLs), stopword lists are often created using the term frequency approach [100, 166], where the top- n most frequent words in a corpus are identified and then manually validated by humans to remove non-stopwords. However, studies have shown that function words typically exhibit high topological connectivity in co-occurrence networks [38, 99, 142], suggesting that network-based approaches may offer a more accurate alternative to traditional frequency-based methods for stopword detection. This hypothesis was confirmed in a collaborative study [10], which demonstrated the greater effectiveness of network-based stopword detection across both low- and high-resource languages.

This chapter reports findings from two original papers developed as part of this thesis: one proposing a novel approach to stopword detection [10] and another presenting the development of a rule-based stemmer for Tetun [73]. These studies are discussed in the following sections.

5.1 Stopword Detection

Stopword lists are essential resources for IR and NLP. While such lists are readily available for high-resource languages, they are often unavailable for LRLs. Manually creating these lists is both labor-intensive and costly, making automated stopword detection a practical alternative.

Various techniques for stopword detection have been studied, including unsupervised frequency-based methods such as term frequency (TF) [17, 55], normalized inverse document frequency (NIDF) [148], inverse document frequency (IDF), term frequency-inverse document frequency (TF-IDF), and term and document frequency (TDF) [91]. Despite these, the high topological connectivity of function words observed in network properties [38, 99, 142] led us to hypothesize that leveraging co-occurrence network features could outperform traditional unsupervised frequency-based methods in stopword detection tasks.

To test this hypothesis, we modeled each pair of sequential terms in the text corpus as connected nodes in a directed graph. A co-occurrence network was then constructed from the corpus, and network-based feature weights were computed for each term. In this setup, function words were expected to receive the highest scores due to their high topological connectivity. This method was compared with the frequency-based term-weighting approaches and evaluated across four languages: Tetun, Emakhuwa,¹ Portuguese and English.

5.1.1 Data Collection and Preparation

While this study was conducted before the development of the Labadain-30k+ dataset, the Tetun corpus was collected from Timor News [59], with additional content scraped from Tatoli using BeautifulSoup. The Emakhuwa corpus consists of data extracted from the Portuguese–Emakhuwa parallel corpus [9], Wikipedia articles,² radio transcripts, and the Emakhuwa translation of the Constitution of the Republic of Mozambique. The Portuguese and English corpora are sourced from WikiCLIR [198],³ which includes Wikipedia articles in 25 languages. A summary of the dataset used in the experiment is provided in Table 5.1.

Table 5.1: Summary of the Dataset. Each document corresponds to a text line separated by a newline character.

Language	Total documents	Vocabulary
Tetun	32,666	22,111
Emakhuwa	52,238	53,880
Portuguese	50,000	57,960
English	50,000	18,369

To prepare the input text corpus for the experiments, the data was preprocessed by converting all text to lowercase, removing punctuation, special characters, and numbers, and subsequently tokenizing into word tokens. The Tetun tokenizer [74] was applied for Tetun, while whitespace tokenization from NLTK was used for Emakhuwa, Portuguese, and English. Due to the noisy nature of Wikipedia data, the Natura dictionary [221] was used to validate Portuguese words, and the English-words dictionary [236] was used for English.

5.1.2 Co-Occurrence Network Properties

The preprocessed data were used to construct co-occurrence networks for each language, where pairs of sequential terms in the corpus were represented as linked nodes in a directed graph, as illustrated in Figure 5.1. Given a directed graph $G = (V, A)$ representing a co-occurrence network, V denotes the nodes, and A represents the edges. Each node $v \in V$ corresponds to a word from the preprocessed vocabulary of the input text document collection, while each adjacency arc

¹A Bantu language primarily spoken in northern and central Mozambique, with approximately 7 million speakers.

²<https://incubator.wikimedia.org/wiki/Category:Wp/vmw>

³<https://www.cs.jhu.edu/~kevinduh/a/wikiclr2018/>

$a \in A$ represents the connection between a pair of consecutive words. The network properties are formulated as follows:

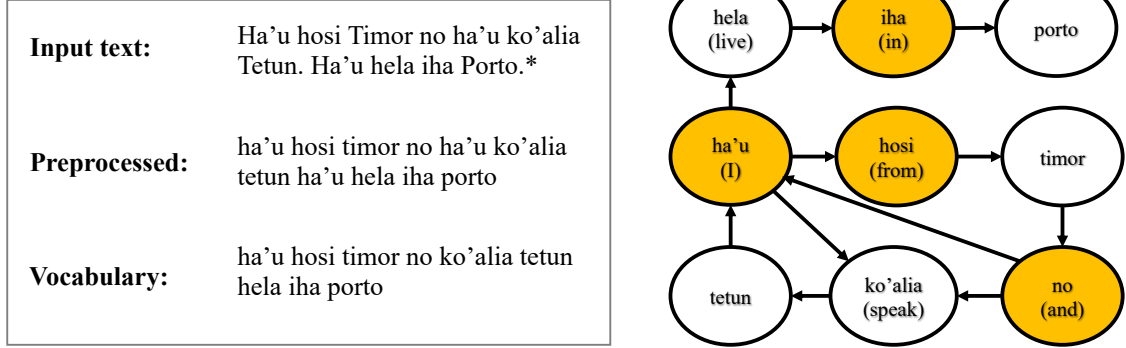


Figure 5.1: Example of a co-occurrence network generated from the preprocessed text. *English translation: “I am from Timor and I speak Tetun. I live in Porto”.

1. In-Degree [88]: The total number of incoming edges to a node v .

$$\text{in-degree}(v) = \sum_{u \in V} A_{uv} \quad (5.1)$$

where A_{uv} is the adjacency matrix element, which equals 1 if there is an edge from node u to v and 0 otherwise.

2. Out-Degree [88]: The total number of outgoing edges from the node v .

$$\text{out-degree}(v) = \sum_{w \in V} A_{vw} \quad (5.2)$$

where A_{vw} is the adjacency matrix element, which equals 1 if there is an edge from node v to w and 0 otherwise.

3. Degree [88]: The sum of a node's in-degree and out-degree, representing its total connections for the node v .

$$\text{degree}(v) = \text{in-degree}(v) + \text{out-degree}(v) \quad (5.3)$$

4. Weighted In-Degree [21]: The sum of the weights of all incoming edges to the node v .

$$\text{weighted in-degree}(v) = \sum_{u \in V} W_{uv} \quad (5.4)$$

where W_{uv} is the weighted adjacency matrix element, representing the weight of the edge from node u to v .

5. Weighted Out-Degree [21]: The sum of the weights of all outgoing edges from the node v .

$$\text{weighted out-degree}(v) = \sum_{w \in V} W_{vw} \quad (5.5)$$

where W_{vw} is the weighted adjacency matrix element, representing the weight of the edge from node v to w .

6. Weighted Degree [21]: The sum of a node's weighted in- and out-degree, representing the total weighted connections for the node v .

$$\text{degree}(v) = \text{weighted in-degree}(v) + \text{weighted out-degree}(v) \quad (5.6)$$

7. Betweenness Centrality [28]: The proportion of shortest paths in the network that pass through the given node v .

$$\text{betweenness centrality}(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \quad (5.7)$$

where σ_{st} is the total number of shortest paths from node s to node t , and $\sigma_{st}(v)$ is the number of those paths that pass through the node v .

5.1.3 Traditional Frequency-Based Methods

Traditional term-weighting approaches frequently employed for stopwords detection include term frequency (TF), normalized term frequency (NTF), inverse document frequency (IDF), normalized inverse document frequency (NIDF), document frequency (DF), term frequency-inverse document frequency (TF-IDF), and term-document frequency (TDF). While some of these techniques have been described in Subsection 2.1.7.1 of Chapter 2, the remaining methods are defined as follows:

1. Normalized term frequency (NTF) [17]: The frequency of a term relative to the total number of terms in the corpus.

$$\text{ntf}(t) = \frac{f(t, d)}{\sum_{t' \in d} f(t', d)} \quad (5.8)$$

where $f(t, d)$ denotes the frequency of term t in document d .

2. Document frequency (DF) [91]: The number of documents in which a term occurs.

$$df(t) = S(t) \quad (5.9)$$

where S is the total number of documents in the corpus containing term t .

3. Term document frequency (TDF) [91]: The product of a term's total occurrences in the corpus and the total number of documents in which it appears.

$$tdf(t) = f(t, d) \times S(t) \quad (5.10)$$

5.1.4 Experiment and Evaluation

The experiment began with feature selection to identify and exclude network attributes that did not significantly contribute to distinguishing stopwords. Each term in the English and Portuguese vocabularies (i.e., nodes) was assigned a Boolean label: `true` if the term was a known stopword and `false` otherwise, using the NLTK stopword list as the reference. Nodes and edges—representing co-occurrence relationships—were then loaded into Gephi⁴ to compute scores for the network properties used as features in subsequent analysis.

To identify the most relevant network features, each network attribute was evaluated based on its information gain relative to the target class (i.e., stopword vs. non-stopword). This analysis was conducted using Weka [96], aiming to reduce the full set of attributes to a smaller, more informative subset that contributes significantly to distinguishing stopwords. Information gain (IG) [201] was computed with respect to the target class using the following formula:

$$IG(S, A) = H(S) - H(S|A) \quad (5.11)$$

Here, $H(S)$ represents the entropy of the entire dataset S , capturing the overall uncertainty before splitting on a given attribute/feature A . Meanwhile, $H(S|A)$ denotes the conditional entropy of S given attribute A , reflecting the remaining uncertainty after the split. The $IG(S, A)$ measures the reduction in entropy achieved by incorporating attribute A , indicating its contribution in predicting the target class. A feature with high information gain significantly reduces entropy, meaning it contributes substantial information to classification. Conversely, a feature with low information gain has minimal impact on entropy reduction and is therefore less effective for distinguishing between classes.

The selected features were then used to compute scores for each term, which were ranked in descending order for attributes such as TF, NTF, DF, betweenness centrality, indegree, outdegree, and degree, or in ascending order for attributes like IDF, NIDF, and TF-IDF. These rankings served as the basis for classifying each term as a stopword or a non-stopword.

For evaluation, the NLTK stopword lists were used to validate stopword detection for Portuguese and English. Since predefined stopword lists for Tetun and Emakhuwa were unavailable, the top 50 terms ranked by their scores were manually translated into English. These translations were cross-referenced with the NLTK English stopword list to ensure accuracy and confirm their relevance as stopwords. The effectiveness of the stopword detection methods was assessed using top- n precision cutoffs.

⁴<https://gephi.org/>

5.1.5 Results and Discussion

Subsection 5.1.5.1 presents and discusses the experimental results for Portuguese and English. Observations on the application of the proposed approaches to Tetun and Emakhuwa are detailed in Subsection 5.1.5.2.

5.1.5.1 English and Portuguese

The importance of network features for stopwords detection tasks is presented in Table 5.2. For English, degree emerged as the top-ranked feature, followed by in-degree. In contrast, for Portuguese, betweenness centrality was the most relevant attribute, with in-degree and out-degree also showing significant importance. Based on these results, the following features were selected for subsequent experiments: betweenness centrality, in-degree, out-degree, and degree.

Table 5.2: Feature ranking with information gain.

Rank	English	Score	Portuguese	Score
1	Degree	0.03525	Betweenness Centrality	0.01223
2	Indegree	0.03515	Indegree	0.01162
3	DF	0.03342	Degree	0.01137
4	NIDF	0.03330	Outdegree	0.01123
5	IDF	0.03330	TF-IDF	0.00988
6	Outdegree	0.03298	Weighted outdegree	0.00985
7	TF	0.03272	TF	0.00983
8	Weighted outdegree	0.03269	NTF	0.00983
9	TF-IDF	0.03252	IDF	0.00982
10	Weighted degree	0.03238	NIDF	0.00982
11	NTF	0.03210	DF	0.00982
12	TDF	0.03170	Weighted Degree	0.00979
13	Betweenness centrality	0.03150	Weighted indegree	0.00973
14	Weighted indegree	0.03038	TDF	0.00969

The results for English are presented in Table 5.3, where the in-degree approach achieved the highest average precision scores among all methods and significantly outperformed the classic TF-IDF approach (paired t-test, $p < 0.01$). Degree performed similarly, as the second-best approach, outperforming all frequency-based methods except at P@175 and P@200. Betweenness centrality yielded results comparable to frequency-based methods, with a slight advantage at P@25. In contrast, out-degree proved less effective for English, achieving a score marginally lower than that of the frequency-based methods.

Table 5.4 presents the results for Portuguese. Network-based approaches demonstrated a clear advantage over frequency-based methods. Betweenness centrality achieved the highest scores across most precision cutoff levels, except at the top 75 cutoff. However, due to its computational complexity, approximately $O(N^3)$ [22], it was the least efficient approach. In contrast, in-degree

Table 5.3: Performance of stopwords detection methods for English.

Approach	P@25	P@50	P@75	P@100	P@125	P@150	P@175	P@200
tf	0.960	0.800	0.706	0.600	0.528	0.480	0.451	0.420
ntf	0.960	0.800	0.706	0.600	0.528	0.480	0.451	0.420
df	0.920	0.800	0.706	0.620	0.544	0.500	0.463	0.410
idf	0.920	0.800	0.706	0.620	0.544	0.500	0.463	0.410
nidf	0.920	0.800	0.706	0.620	0.544	0.500	0.463	0.410
tf-idf	0.920	0.800	0.693	0.610	0.560	0.500	0.457	0.415
tdf	0.960	0.800	0.693	0.620	0.552	0.500	0.457	0.415
betweenness centrality	1.000	0.800	0.693	0.620	0.552	0.487	0.429	0.390
indegree	1.000	0.880	0.800	0.690	0.608	0.540	0.469	0.415
outdegree	0.920	0.780	0.667	0.570	0.504	0.467	0.423	0.380
degree	1.000	0.840	0.800	0.700	0.608	0.520	0.460	0.411

and degree yielded results comparable to betweenness centrality, with degree slightly outperforming in-degree. This can be attributed to the degree being the summation of in-degree and out-degree, making it inherently proportional to both. Similar to the English results, out-degree was the weakest performer among the network-based approaches for Portuguese. Again, in-degree outperformed the TF-IDF with statistical significance (paired t-test, $p < 0.01$).

Table 5.4: Performance of stopwords detection methods for Portuguese.

Approach	P@25	P@50	P@75	P@100	P@125	P@150	P@175	P@200
tf	1.000	0.680	0.533	0.490	0.432	0.373	0.331	0.310
ntf	1.000	0.680	0.533	0.490	0.432	0.373	0.331	0.310
df	0.920	0.680	0.560	0.520	0.448	0.386	0.337	0.310
idf	0.920	0.680	0.560	0.520	0.448	0.386	0.337	0.310
nidf	0.920	0.680	0.560	0.520	0.448	0.386	0.337	0.310
tf-idf	0.920	0.680	0.546	0.520	0.448	0.373	0.337	0.315
tdf	0.960	0.680	0.546	0.510	0.440	0.373	0.331	0.310
betweenness centrality	1.000	0.920	0.747	0.620	0.552	0.480	0.440	0.400
indegree	1.000	0.860	0.733	0.600	0.528	0.460	0.417	0.375
outdegree	0.960	0.800	0.707	0.590	0.512	0.447	0.411	0.370
degree	1.000	0.900	0.773	0.620	0.528	0.467	0.406	0.380

Our findings align with the claims of Gao et al. [99], Chen et al. [38] and Liang et al. [142], who argue that stopwords are highly connected nodes within a co-occurrence network. Additionally, our results suggest that degree and in-degree are more effective for stopwords detection compared to frequency-based methods.

5.1.5.2 Tetun and Emakhuwa

To further evaluate the effectiveness of the network-based approaches for stopwords detection, experiments were conducted on Tetun and Emakhuwa. The results are summarized in Table 5.5. Due to the additional manual effort required to project (or translate) stopwords from English to

Tetun and Emakhuwa, the analysis focused on the top 50 ranked words. The results demonstrate that the network-based approaches consistently outperformed frequency-based methods, with in-degree achieving the highest performance for both Tetun and Emakhuwa.

Table 5.5: Performance of stopwords detection methods for Tetun and Emakhuwa

Approach	Tetun		Emakhuwa	
	P@25	P@50	P@25	P@50
tf	0.840	0.720	0.760	0.740
ntf	0.840	0.720	0.760	0.740
idf	0.880	0.760	0.800	0.760
nidf	0.640	0.500	0.800	0.760
tf-idf	0.880	0.760	0.800	0.760
tdf	0.880	0.740	0.760	0.760
betweenness centrality	0.920	0.820	0.720	0.780
indegree	1.000	0.900	0.760	0.820
outdegree	0.960	0.840	0.760	0.800
degree	1.000	0.860	0.760	0.800

The top 25 stopwords for Tetun and Emakhuwa are visualized in Figure 5.2 and 5.3, respectively. The networks were filtered by in-degree and partitioned into stopwords (green nodes) and non-stopwords (pink nodes). Each node is labeled with the original word and its English translation, separated by a colon.

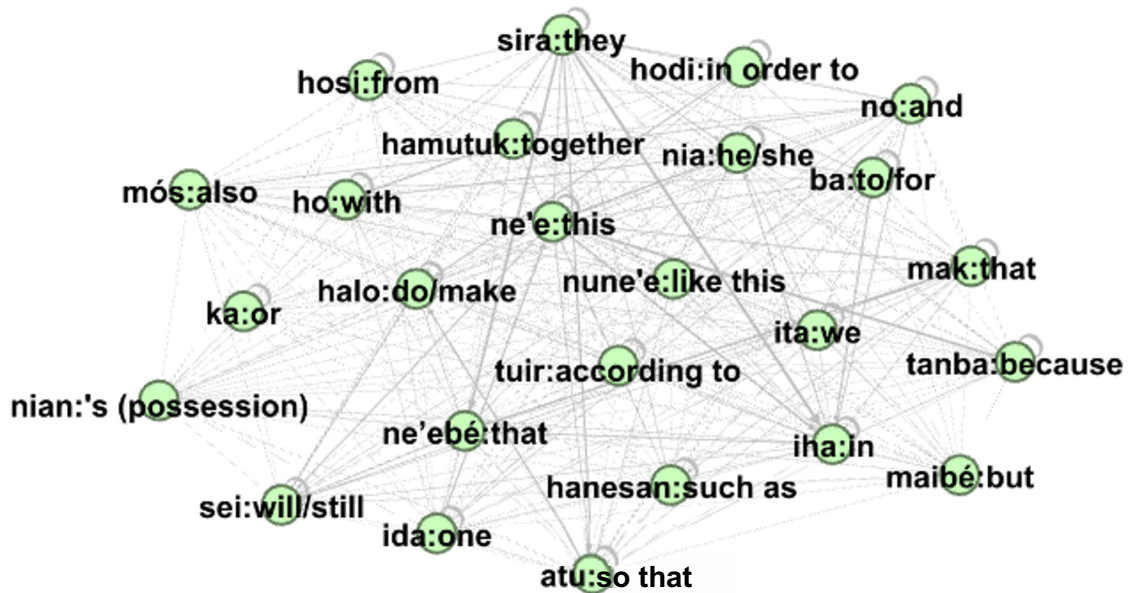


Figure 5.2: Top-25 stopwords in Tetun identified using in-degree.

For Emakhuwa, a noticeable precision drop is observed in Figure 5.3, with five misclassified words. This reduction in precision is likely due to the nature of the Emakhuwa corpus, which predominantly consists of religious texts. Such texts often contain a high frequency of religiously

themed words, such as “*Yesu*” (meaning “Jesus”), “*Muluku*”/“*Yehova*” (meaning “God”), and others that do not align with conventional stopword definitions.

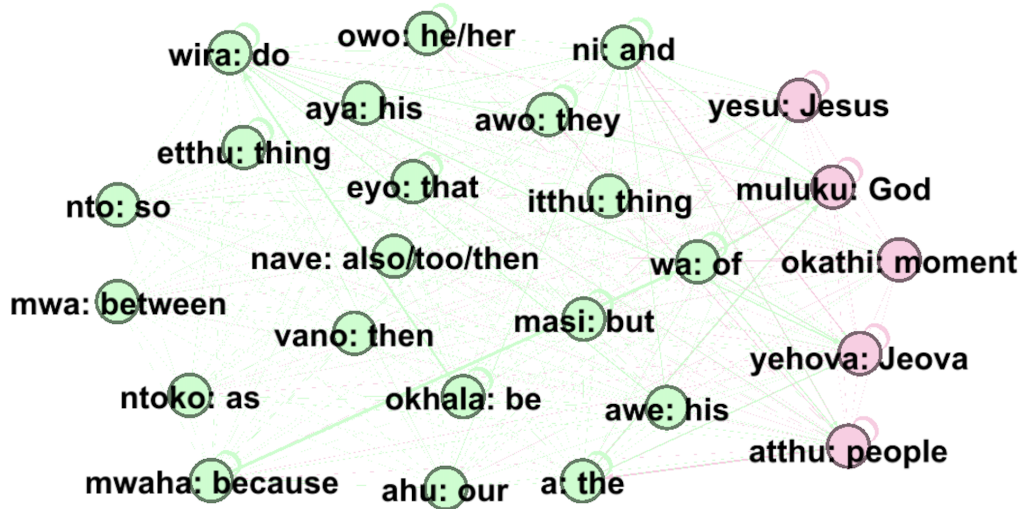


Figure 5.3: Top-25 stopwords in Emakhuwa identified using in-degree.

5.1.6 Conclusions

This study proposes an automated approach for stopword detection, leveraging network properties extracted from sequential term pairs within text corpora. The datasets were pre-processed to construct term-pair co-occurrence networks, and relevant network features were selected based on their information gain scores. These features were then evaluated across four languages: Tetun, Emakhuwa, Portuguese, and English. Experimental results show that network-based features consistently outperform traditional frequency-based methods, with in-degree emerging as the most effective and reliable feature for identifying stopwords.

5.2 Tetun Stopword List Building

Network-based approaches have demonstrated superior performance over traditional frequency-based methods in stopword detection tasks. In previous work [10], a Tetun stopword list was manually created by translating the top 50 terms. Building on this foundation, this study aims to validate the proposed methods by systematically constructing a stopword list for Tetun and using it to evaluate the effectiveness of the top-performing methods identified earlier.

This section outlines the process of developing *Labadain-Stopwords*, a Tetun stopword list. It describes the methodology for constructing the list and evaluates the effectiveness of network-based methods compared to traditional frequency-based techniques, using datasets in English, Portuguese, and Tetun, which differ from those used in the previous study.

5.2.1 Methodology

The *Labadain-Stopwords* list was built using Labadain-30k+ [64]. The process began with preprocessing steps, including lowercasing, normalizing apostrophes, and removing punctuation, special characters, numbers, and extra spaces, followed by tokenization with the Tetun tokenizer [74] and deduplication to create a vocabulary. Traditional term-weighting techniques—TF, IDF, and TF-IDF—were then applied to the vocabulary to assign weights to each word in the vocabulary.

To analyze network properties, a vocabulary-level co-occurrence network was built as a directed graph from the preprocessed text, where each word was represented as a node. The graph properties were analyzed by computing in-degree (total incoming links), out-degree (total outgoing links), and degree (sum of incoming and outgoing links) for each node, reflecting the connectivity of each word within the network.

Using these weights, the top 1,000 words from each method were selected in ascending or descending order based on their scores (as explained in Subsection 5.1.4) to generate lists of potential stopwords. These lists were merged, with duplicates and misspelled words eliminated, to produce a consolidated list of candidate stopwords. Two native Tetun speakers, a Ph.D. student and an undergraduate student, reviewed and validated this list, resulting in the final *Labadain-Stopwords*, which contains 160 Tetun stopwords. The complete list, along with English translations, is provided in Section A.4 of Appendix A.

Some stopwords appeared in misspelled forms. For example, the word “*ne’ebé*” (meaning “which/that” in English) was found in variations such as “*nebe*”, “*neebe*”, and “*neebé*”. These variations were compiled into a separate list of stopwords variations, which was subsequently used to develop a stopword correction list.

5.2.2 Experiment and Evaluation

In the previous study [10], experiments were conducted using a limited Tetun text corpus collected from two news portals, and the Tetun stopword list was manually translated from English to serve as ground truth. In this study, we use the Labadain-30k+ dataset [64], with the curated *Labadain-Stopwords* list for evaluating the effectiveness of network-based stopword detection approaches. To further assess the method’s generalizability across both low- and high-resource languages, additional experiments were carried out using datasets in English and Portuguese.

For the Portuguese and English datasets, the documents used in the experiments were extracted from the CC-100 dataset [235], and stopword lists from NLTK served as ground truth. The process of computing weights for Portuguese and English words followed the same methodology applied to Tetun, ensuring consistency across experiments. A summary of the datasets used for generating the stopword lists is presented in Table 5.6.

For evaluation, while previous work limited the analysis to P@200, this study extended the P@*n* cutoff to 1,000. Cutoffs were applied at intervals of approximately 25 for positions up to 100 and at intervals of 250 for positions between 100 and 1,000.

Table 5.6: Summary of Tetun, Portuguese, and English datasets used in stopwords detection.

Description	Tetun	Portuguese	English
Number of documents	33,550	3,153	624
Number of words	11,928,821	613,736	667,584
Vocabulary size	146,783	45,860	31,390

5.2.3 Results

The results of the experiment with Tetun, presented in Table 5.7, demonstrate that network-based approaches generally outperform traditional term-weighting methods in identifying stopwords. Specifically, in-degree consistently demonstrates superior performance across most cutoffs, except at P@75, where the degree slightly surpasses it. At P@10, all methods attain perfect precision. At P@25 and P@1000, both in-degree and degree achieve identical performance scores. Among the frequency-based baselines, the results are closely identical, with IDF slightly outperforming TF and TF-IDF at the P@500 and P@1000 cutoffs.

Table 5.7: Stopword precision for Tetun.

Approach	P@10	P@25	P@50	P@75	P@100	P@250	P@500	P@750	P@1000
In-degree	1.0000	0.9600	0.8400	0.7200	0.7000	0.4720	0.3080	0.2347	0.1930
Out-degree	1.0000	0.8800	0.8000	0.6933	0.6000	0.4240	0.2900	0.2160	0.1780
Degree	1.0000	0.9600	0.8200	0.7333	0.6500	0.4640	0.3000	0.2253	0.1930
TF	1.0000	0.9200	0.6800	0.6000	0.5200	0.3600	0.2480	0.1933	0.1610
IDF	1.0000	0.9200	0.6800	0.6000	0.5200	0.3600	0.2540	0.1973	0.1640
TF-IDF	1.0000	0.9200	0.6800	0.5867	0.5100	0.3560	0.2500	0.1947	0.1620

For Portuguese, similar patterns were observed, with network-based methods again demonstrating superior performance, as presented in Table 5.8. At P@10, all methods performed equally well, achieving perfect precision. The degree slightly outperformed in-degree at P@75 and P@100. At P@25 and P@750, both in-degree and out-degree achieve identical scores. In-degree outperforms degree at P@250 and P@500. At P@1000, out-degree outperformed all other approaches.

Table 5.8: Stopword precision for Portuguese.

Approach	P@10	P@25	P@50	P@75	P@100	P@250	P@500	P@750	P@1000
In-degree	1.0000	1.0000	0.9000	0.7600	0.6700	0.3800	0.2260	0.1587	0.1260
Out-degree	1.0000	0.9600	0.9600	0.8000	0.6700	0.3560	0.2180	0.1573	0.1270
Degree	1.0000	1.0000	0.9400	0.8133	0.7100	0.3680	0.2240	0.1587	0.1240
TF	1.0000	0.9600	0.9000	0.7600	0.6500	0.3320	0.2000	0.1480	0.1200
IDF	1.0000	0.9600	0.9000	0.7600	0.6500	0.3320	0.2000	0.1480	0.1200
TF-IDF	1.0000	0.9600	0.9000	0.7600	0.6500	0.3320	0.2000	0.1480	0.1200

Notably, traditional term-weighting approaches produced consistent results across all evaluated cutoffs. To investigate these identical scores, we tested reduced dataset sizes, which revealed slight variations across cutoffs. This suggests that the uniform scores are likely due to inherent characteristics of the dataset.

Table 5.9: Stopword precision for English.

Approach	P@10	P@25	P@50	P@75	P@100	P@250	P@500	P@750	P@1000
In-degree	1.0000	1.0000	0.9400	0.9333	0.8200	0.4120	0.2280	0.1573	0.1220
Out-degree	1.0000	1.0000	0.9800	0.8400	0.7600	0.4160	0.2280	0.1573	0.1220
Degree	1.0000	1.0000	0.9800	0.9067	0.8200	0.4280	0.2300	0.1613	0.1220
TF	1.0000	1.0000	0.9600	0.8933	0.7500	0.4240	0.2280	0.1573	0.1220
IDF	1.0000	1.0000	0.9600	0.8933	0.7500	0.4240	0.2280	0.1573	0.1220
TF-IDF	1.0000	1.0000	0.9600	0.8933	0.7500	0.4240	0.2280	0.1573	0.1220

Similar results to those observed in the Portuguese dataset were noted in English, where network-based approaches maintained their advantage, as shown in Table 5.9. At P@10, P@25, P@100, and P@1000, both in-degree and out-degree achieve identical scores. Degree slightly outperforms in-degree at P@50, P@250, P@500, and P@750. In-degree outperforms degree at P@75. At P@10 and P@25, all techniques achieve perfect precision. At P@1000, all methods achieve identical performance scores. Consistent with the Portuguese dataset, traditional term weighting methods for English produced identical results across all evaluated cutoffs.

5.2.4 Discussion

To analyze the performance of stopwords detection methods across different ranking levels, precision was examined at three cutoff ranges: lower (up to P@25), mid-range (P@50 to P@100), and higher (P@250 to P@750). At the lower cutoffs, all methods produced comparable results, with network-based approaches—particularly in-degree and degree—slightly outperforming traditional term-weighting methods. Specifically, in Tetun and Portuguese, network-based methods achieve a modest gain of +0.04 points at P@25 over their frequency-based counterparts.

At the mid-range cutoffs, network-based methods maintain a consistent advantage, outperforming traditional approaches by up to +0.06 points in Portuguese, +0.07 points in English, and +0.18 points in Tetun. Within this range, degree consistently outperforms in-degree.

At the higher cutoffs, network-based methods continue to outperform traditional frequency-based approaches. In-degree consistently delivers the best results for Tetun and Portuguese, with improvements of up to +0.18 points. In contrast, for English, the degree marginally outperforms in-degree. These findings suggest that network-based methods maintain a stronger advantage as the stopwords list expands, except in English at P@1000, where all methods produce identical scores, likely due to characteristics specific to the dataset.

Overall, in Tetun, in-degree proved slightly more effective than degree, while in English, degree marginally outperformed in-degree. For Portuguese, both methods exhibited similar performance (see Figure 5.4). Given that the degree is the sum of in-degree and out-degree, and its performance is comparable, in-degree offers a computationally more efficient advantage.

For traditional term-weighting approaches, IDF outperformed TF and TF-IDF at higher cutoffs in Tetun. At mid-range cutoffs, IDF and TF achieved identical scores, while at lower cutoffs, TF, IDF, and TF-IDF all produced identical scores. In contrast, for Portuguese and English, TF, IDF,

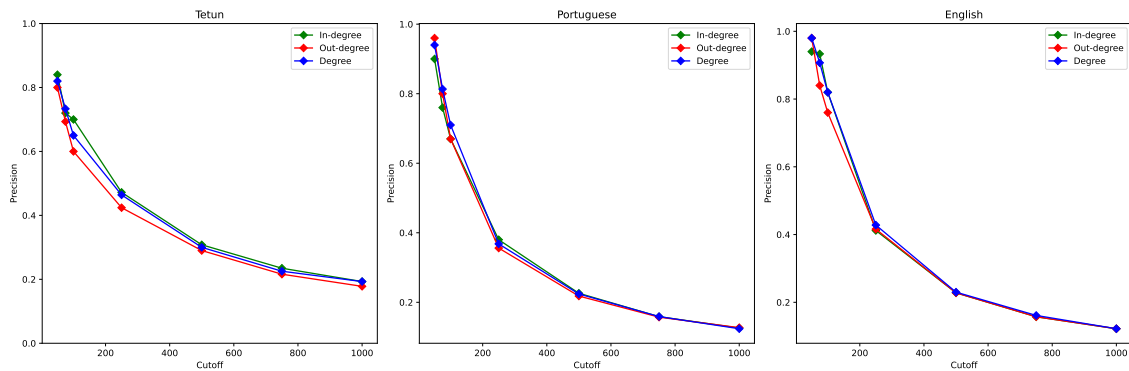


Figure 5.4: Performance of Network-based approaches at mid-range and higher cutoff levels.

and TF-IDF yielded identical results across all cutoffs. This consistency in Portuguese and English may be due to the more structured and mature linguistic resources available for these languages, such as well-established stopwords lists and corpora, which minimize variations in term weighting effectiveness. In Tetun, the language’s lesser-resourced nature likely results in greater sensitivity to different weighting methods, leading to performance differences at higher cutoffs.

Furthermore, the total number of stopwords we developed for Tetun is comparable to other LRLs such as Marathi (99 stopwords), Bengali (114 stopwords), and Hindi (165 stopwords) [192]; and Kinyarwanda (80 stopwords) and Kirundi (59 stopwords) [166].

5.2.5 Conclusions

This study highlights the superiority of network-based approaches, particularly in-degree and degree, over traditional term weighting methods for stopwords detection in both high- and low-resource languages, especially when dealing with larger stopwords lists. While traditional term weighting and network-based methods perform similarly at lower cutoffs (up to 25 words), network-based approaches demonstrate greater effectiveness as the number of evaluated words increases.

For smaller stopwords lists, differences between the methods are less pronounced. However, when the list exceeds 25 stopwords, network-based approaches are recommended due to their superior performance at mid- and higher-range cutoffs. The consistent performance of in-degree, particularly in Tetun, aligns with findings from our previous study [10], further validating the effectiveness of network-based methods for stopwords detection.

These findings underscore both practical applications, such as method selection and scalability, and theoretical advancements, including the validation of network-based methods. This work benefits researchers in IR and NLP, particularly those working with LRLs.

5.3 Labadain-Stemmer Development

Stemming is a text normalization technique in IR, used to reduce words to their base or root form, typically by removing suffixes. This section outlines the development of the *Labadain-Stemmer*,

a stemming algorithm designed for Tetun. It covers the identification of Tetun affixes and the creation of multiple stemmer variants tailored to the language’s characteristics.

The section also describes the construction of a word sample validated by human annotators to establish a ground truth for evaluation. Finally, it presents the experimental setup, results, and key findings, followed by the study’s limitations and main observations.

5.3.1 Tetun Affixes

This study focuses on the commonly used affixes in Tetun [103, 116, 224], excluding circumfixes and reduplication due to their limited productivity and usage, as discussed in Subsection 2.2.4 of Chapter 2. The native Tetun prefixes are “*ha*”, “*nak*”, and “*nam*”, while the native suffixes consist of “*n*”, “*-nain*”, “*-teen*”, and “*dór*”.

Portuguese-derived suffixes were adapted from the Portuguese suffix list used in the Portuguese stemmer in Snowball [206], as presented in Table A.2 of Appendix A. Given that Tetun contains relatively few inflectional affixes, stemming native Tetun words is a straightforward process, employing the longest matching technique [150] on each input word using a predefined list of Tetun prefixes and suffixes.

5.3.2 Labadain-Stemmer Variants

Tetun consists of both native words and a substantial number of Portuguese loanwords, particularly verbs, nouns, and adjectives [103, 106, 223, 224]. To handle this linguistic diversity, the *Labadain-stemmer* is designed with three variants: *light*, *moderate*, and *heavy*. Each of these is described in detail in the following subsections.

5.3.2.1 Light Stemmer

The *light* stemmer is designed to remove suffixes from Portuguese-derived words used in Tetun. This variant adapts the Portuguese stemmer from Snowball, incorporating a customized list of Portuguese suffixes. These suffixes were modified based on the loanword transformation rules defined by the INL [168], as detailed in Table 5.10.

Table 5.10: Transformation rules for Portuguese-derived words applied to Tetun suffixes.

Rule (Portuguese → Tetun)	Effect on suffix	Example (Portuguese → Tetun)
ão → aun	asaun	comemoração → komemorasaun (celebration)
ss, c (before e, i), ç (before a, o, u) → s	saun	discussão → diskusaun (discussion)
qu, c (before a, o, u) → k	ik + amente	automaticamente → automatikamente (automatically)
g (before e, i) → j	lojia	tecnologia → teknolojia (technology)
s (between vowels) → z	oza	poderosa → poderoza (powerful)
ê → é	énsia	competência → kompeténsia (competence)
â → á	ánsia	ignorância → ignoránsia (ignorance)
o → u	u	infermeiro → infermeiru (nurse)

The Tetun *light* stemmer is a simplified adaptation of the Portuguese stemmer, designed to handle loanwords while accounting for Tetun’s unique morphological characteristics. It retains the linguistic regions utilized in the original Portuguese stemmer, which were adapted from the Spanish stemmer in Snowball [205]. The definitions of these linguistic regions, as applied in the Tetun *light* stemmer algorithm, are provided in Table A.1 of Appendix A.

The Tetun *light* stemmer processes words sequentially using a list of suffixes developed to account for the specific features of Portuguese loanwords in Tetun (see Table A.2). The stemming procedure is summarized below, with the corresponding algorithm provided in Algorithm 1:

1. *Word length validation*: After receiving an input word, the algorithm begins by validating its length. If the word contains fewer than four characters, it is returned without stemming.
2. *Standard suffix removal*: For words longer than three characters, the algorithm searches for the longest matching suffix from the general suffix list. If a matching suffix is found in a specific region of the word, the suffix is deleted or replaced accordingly.
3. *Verb suffix removal*: If no suffix is removed in step (2), the algorithm checks for verb-specific suffixes. It is removed if a matching suffix is found within the appropriate region of the word.
4. *Residual suffix removal*: If neither of the previous steps results in suffix removal, the algorithm looks at the remaining simple suffixes list and removes it as the final step.
5. *Return original word*: If none of the steps result in suffix removal, the input word is returned unchanged.

5.3.2.2 Moderate Stemmer

The *moderate* stemmer extends the functionality of the *light* stemmer by handling suffixes from both Portuguese loanwords and native Tetun. It adheres to the same algorithm as the *light* stemmer (outlined in Algorithm 1), with the addition of a new step of 4.1, specifically designed to process native Tetun suffixes. This additional step is executed between steps 4 and 5 of the algorithm.

5.3.2.3 Heavy Stemmer

The *heavy* stemmer builds on the functionality of the *moderate* stemmer by introducing the removal of native Tetun prefixes. In this variant, the processing of native Tetun prefixes is integrated between steps 4.1 and 5 of the algorithm.

5.3.3 Text Sample Construction for Evaluation

Selecting sample words to evaluate stemming performance presents challenges, such as potential bias and limited generalizability. To mitigate these issues, we developed a systematic methodology to construct the sample from a dataset encompassing a diverse range of categories and sources and

incorporating human involvement in the process. This subsection describes the creation of the sample words used to evaluate the *Labadain-Stemmer* for intrinsic assessment purposes.

The process of constructing the sample words is presented in Figure 5.5. First, the Labadain-30k+ dataset [64] was preprocessed by converting the text to lowercase, normalizing apostrophes, and removing punctuation, special characters, numbers, and extra spaces. This preprocessed text was then tokenized into individual words using the Tetun Word Tokenizer [67], and deduplication was performed to remove duplicate tokens. Stopwords were normalized and subsequently excluded, along with tokens shorter than four characters. To further refine the candidate sample words, the Tetun LID model [66] was applied with a threshold score of 0.95 to filter out words that did not meet this criterion. Finally, the refined candidate sample words were validated by native Tetun speakers to produce the final set of sample words.

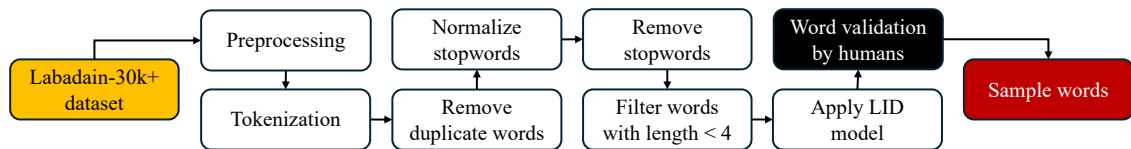


Figure 5.5: Process of constructing a text sample for evaluating the Tetun stemmer’s performance.

The sample word verification with human involvement was carried out by six native Tetun-speaking students, consisting of one Ph.D. student and five undergraduate students. Each student was tasked with verifying approximately 2,732 words, sorted in ascending alphabetical order. They checked the correctness of each word using both the INL dictionary [49] and the Portuguese loanword dictionary [103] as reference materials.

During the verification process, a considerable number of misspelled words were identified, such as the word “*konsiderasaun*” (consideration, in English) appearing as [*konsidersaun*, *konsiderasaunn*, *konsideransaun*]. Additionally, some words originating from Tetun Terik, Tetun DIT, or other variants not present in the reference dictionaries were excluded from the final sample. A summary of the resulting sample of words from this process is presented in Table 5.11.

Table 5.11: Summary of words produced at each stage of the sample word generation process. *Approximately 88.4% reduction in the number of words was observed after applying the LID model, attributed to the model’s average score per word of 0.394, with a threshold set to 0.95.

Description	#words
Initial total number of words	146,387
Remaining words after removing stopwords	146,204
Remaining words after excluding words with fewer than four characters	141,487
Remaining words after applying the LID model	16,391*
Remaining words after manual verification by human assessors (final sample words)	1,839

5.3.4 Ground Truth Development

The development of the ground truth set for evaluating the *Labadain-Stemmer* performance was carried out by the same six Timorese students. To familiarize the assessors with the evaluation process, five example pairs from the sample of 1,839 words (see Table 5.11) were provided during the training session. These pairs included the original words and their corresponding stemmed forms generated by each *Labadain-Stemmer* variant.

Following the training session, the complete list of input words and their corresponding stemmed results was distributed to two students per stemmer variant for evaluation. Their primary task was to assess whether each word correctly stemmed to its root form. In cases where a word was incorrectly stemmed, the students provided the correct root form, using the suffixes outlined in Table A.2 for Portuguese-derived words and the Tetun affixes described in Subsection 5.3.1 to guide their decisions.

Inter-annotator agreement was calculated to ensure consistency and reliability among the annotators. Discrepancies between annotators were analyzed and discussed, allowing them to reach a consensus for each stemmer variant. All annotators followed this procedure during the evaluation of the different stemmer variants. Inter-annotator agreement was measured using Cohen’s kappa, as presented in Table 5.12.

Table 5.12: Cohen’s k score for inter-annotator agreement in constructing the ground truth set.

Algorithm	k -Score
Light stemmer	0.7006
Moderate stemmer	0.6990
Heavy stemmer	0.7683

In the final stage, the annotators consolidated their evaluations and resolved any remaining discrepancies to finalize the correct stemmed forms. These consensus-based results were then used to compile the ground truth set, which is summarized in Table 5.13.

Table 5.13: Summary of the ground truth set and word length.

Description	All	PT loanwords	Native Tetun
Total number of words	1,839	81.79%	18.21%
Minimum character count per word	4.00	4.00	4.00
Maximum character count per word	20.00	20.00	15.00
Average character count per word	9.50	10.21	5.97

5.3.5 Evaluation Metric

The Paice metric [174] was used to evaluate the stemmer variant quality by measuring how effectively they reduce various word forms to a common root. This metric balances understemming and overstemming effects, both of which impact precision and recall in text-processing tasks. In

IR, a high understemming lowers recall, resulting in relevant documents not being retrieved, while a high overstemming hurts precision by retrieving many irrelevant documents.

Paice introduced four intrinsic methods to assess stemming performance: understemming index (UI), overstemming index (OI), stemming weight (SW), and error rate relative to truncation (ERRT). UI measures how often the stemmer fails to reduce related words to the same root, while OI calculates the frequency of incorrectly merging unrelated words into the same root. SW is the ratio of OI/UI, representing the trade-off between overstemming and understemming. A lower value of SW indicates more understemming, whereas a higher value suggests a tendency toward overstemming. ERRT evaluates the stemmer's ability to balance understemming and overstemming. This involves computing UI and OI values for various truncation lengths to establish a *truncation line*, which serves as a baseline for stemmer performance. Any reasonable stemmer should have its (UI, OI) point located between the truncation line and the origin, with better performance indicated by a position further away from the truncation line or closer to the origin.

5.3.6 Experimental Setting

To compute the Paice metric, a list of words is first organized into conceptual groups based on semantic and morphological relationships. These groups serve as the target, and an ideal stemmer should conflate words according to these conceptual groupings. The stemmers were then applied to the word list, and their performance was evaluated by measuring how accurately they matched the predefined conceptual groups. Examples of these conceptual groupings are provided below, where the root word is shown on the left side and its corresponding conflated words are listed on the right side, separated by a colon delimiter.

```
"ajente": ["ajénsia", "ajénsias"]
"akompañ": ["akompaña", "akompañamentu", "akompañante"]
"akontes": ["akontese", "akontesimentu", "akontesimentus"]
"hatete": ["hatete", "hateten"]
"kbiit": ["kbiit-laek", "kbiit-na'in", "kbiit"]
"komunik": ["komunikadu", "komunikasaun", "komunikativa"]
"otél": ["otél"]
```

The application of the stemmer to conceptual groups resulted in understemming, overstemming, and the relative accuracy of the stemmers, represented by ERRT. To calculate ERRT, a baseline was established using length truncation, where the words in the list were truncated to their first n letters, with n set to 7, 8, and 9. The overstemming and understemming measures of these truncated lists define the truncation line.

5.3.7 Evaluation and Results

Using the Paice metric, we calculated the ERRT value for each stemmer variant by drawing a line from the origin through the point representing its understemming and overstemming indexes

(UI, OI) and extending it to intersect the truncation line. The ERRT is calculated by dividing the distance from the origin to the (UI, OI) point by the distance from the origin to the truncation line intersection. An ideal stemmer variant has low UI and OI values, indicating better performance when positioned closer to the origin or further away from the truncation line. Figure 5.6 presents the UI and OI values for each stemmer with the truncation line, showing that the *heavy* and *moderate* stemmer variants slightly outperformed the *light* variant.

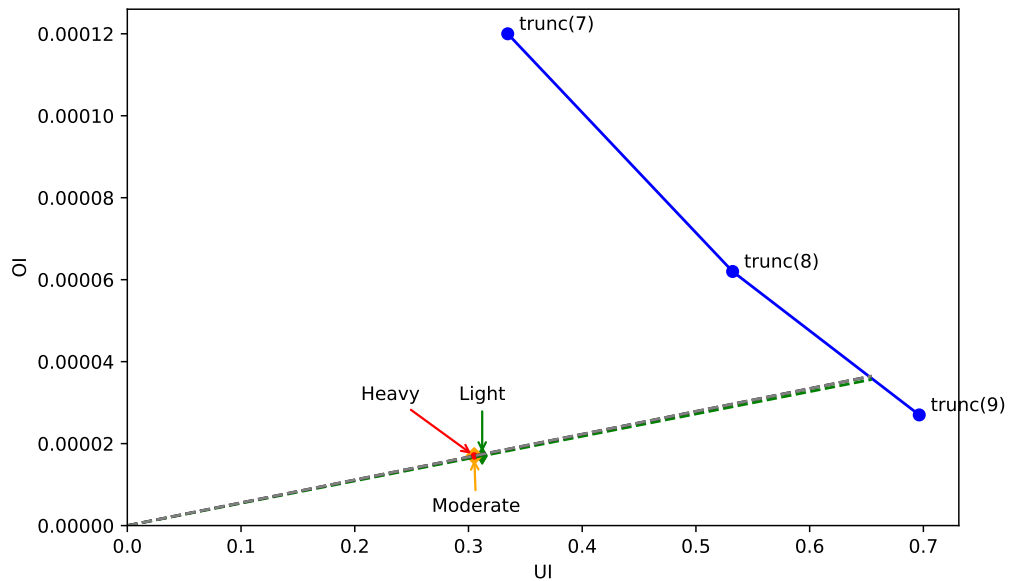


Figure 5.6: UI vs. OI plot showing ERRT distances.

The UI, OI, SW, and ERRT values are presented in Table 5.14. As expected, the *light* variant exhibits the highest understemming value (by +0.007 points), while the *moderate* and *heavy* variants have identical lowest ERRT values (both lower by approximately -0.009 points), and all variants have the same overstemming values. Further investigation revealed that the difference in words stemmed from the *heavy* variant compared to the *moderate* variant was limited to only six words. This is due to the small proportion of native Tetun words, which make up only 18.21% of the total (see Table 5.13), and the limited number of Tetun prefixes (outlined in Subsection 5.3.1).

Table 5.14: Analysis of the stemming algorithms' performance using the Paice metric.

	UI	OI	SW	ERRT
Light	0.312062	0.000017	0.000056	0.481367
Moderate	0.305049	0.000017	0.000057	0.472802
Heavy	0.305049	0.000017	0.000057	0.472802

In terms of overstemming value, since native Tetun has only a few inflectional forms in the word list, applying both the *moderate* and *heavy* variants had no impact on the overstemming value. The minimal difference in the ERRT values presented in Table 5.14 indicates that the *light*, *moderate*, and *heavy* stemmer variants performed quite similarly.

5.3.8 Discussion

Tetun, as a language with relatively few inflectional affixes [103, 116, 224], often includes short affixes, such as the prefix “*ha*” and the suffix “*n*”, which present challenges to stemming algorithms in correctly handling native Tetun affixes. Some verbs and nouns begin or end with these characters, although they are not affixes, for instance in words like “*halimar*” (play), “*hariis*” (bathe), “*aman*” (father), “*inan*” (mother), “*ibun*” (mouth), “*liman*” (hand), “*ulun*” (head), among others. Additionally, removing these characters from certain words changes their meaning. For example, removing the prefix “*ha*” from “*halimar*” (play) results in “*limar*” (rasp), and the suffix “*n*” from “*liman*” (hand) becomes “*lima*” (five).

Given that three variants of the *Labadain-Stemmer* show similar performance, it might be affected by the characteristics of the sample words used for evaluation. Factors such as the proportion of native Tetun words, the presence of affixes, and the term distribution within the sample could impact the stemming algorithms’ outcomes. Developing a more balanced sample word composition could provide deeper insight into the observed results.

5.3.9 Limitations

The *Instituto Nacional de Linguística* (INL) launched the *Kursu Gramátika Tetun* (Tetun Grammar Course) in 2005, which has served as the key reference for teachers, translators, journalists, and students [116]. The book covers several Tetun affixes, including prefixes such as [*hak*, *na*, *ma*] and suffixes like [*-laek*, *-k*]. However, more recent research by Greksáková [103] in 2018 reports that many of these affixes have been largely replaced by words such as “*sai*” (become) and “*laiha*” (without), making these affixes less productive in Tetun. Since the INL has not updated its 2005 publication on Tetun grammar, the current state of Tetun morphology remains uncertain.

Additionally, the lack of linguistic expert involvement, due to limited funding, is a significant limitation of this study. Nevertheless, the baseline established here provides a foundation for future research on Tetun.

5.3.10 Conclusions

This study developed and evaluated the effectiveness of the *Labadain-Stemmer*, which incorporates suffixes from Portuguese loanwords and affixes from native Tetun. The affixes were selected based on the commonly used Tetun affixes reported by van Klinken et al. [224], the INL [116], and Greksáková [103]. To assess the stemmer’s performance, we systematically constructed sample words and established a baseline for the *Labadain-Stemmer*, testing three variants: *light*, *moderate*, and *heavy*. The results showed that integrating native Tetun affixes into the stemming process was slightly more effective than focusing exclusively on the suffixes of Portuguese loanwords.

However, one of the limitations in this study is the unbalanced representation of Portuguese loanwords and native Tetun words in the sample set. Future research should address this by employing more balanced sample words that adequately reflect both Portuguese loanwords and native Tetun vocabulary. Additionally, involving expert linguists specializing in Tetun will be essential

for further investigating the accuracy and overall effectiveness of the *Labadain-Stemmer*. To enable reproducibility, the stemmer algorithms have been released under the MIT License [75], to encourage further research and development in Tetun information retrieval.

The findings of this study make a significant contribution to advancing IR in Tetun by providing a replicable methodology for developing stemming algorithms tailored to LRLs influenced by loanwords. Additionally, the construction of a ground truth set for evaluating stemmer performance offers a structured and adaptable approach that supports the development of effective stemmers for other low-resource languages.

Chapter 6

Tetun Test Collection

The advancement of information retrieval (IR) systems depends on reliable test collections to evaluate their effectiveness. Prior to this work, no such collection existed for Tetun, making its development a crucial step forward. This chapter reports on the construction of a Tetun test collection following TREC guidelines and explores the feasibility of using large language models (LLMs) to automate this process in low-resource language (LRL) scenarios.

This effort resulted in the publication of two original papers: one established a foundation for Tetun ad-hoc text retrieval [73], and the other explored the use of LLMs for relevance judgments in Tetun [65]. Both studies are presented in this chapter.

6.1 Overview

The classical approach to constructing test collections follows the Cranfield paradigm [41], which gained widespread recognition through the TREC series of large-scale evaluation campaigns [107]. In this TREC-style adaptation of the Cranfield approach, a test collection comprises three components: a document collection, a set of information needs (or *topics*), and corresponding relevance judgments. In ad-hoc text retrieval, a set of topics is formulated and then tested by searching a large document collection to estimate the number of relevant documents returned for each topic [194].

These query-document pairs are then provided to assessors for relevance judgments. Traditionally, relevance judgments are made by human assessors, a process that is both labor-intensive and costly. Due to financial constraints, relevance assessment tasks for constructing test collections for LRLs are often carried out by volunteer native language speakers, such as students [8, 191].

To mitigate the challenges of manual relevance judgments, the IR community has been investigating the feasibility of automating this task. With the advent of LLMs, which have demonstrated strong performance across various tasks, new possibilities for conducting automated relevance judgments have emerged. These models have demonstrated ongoing improvement in the quality of automated relevance judgment tasks as LLMs continue to evolve, offering promising alternatives for contributing to the construction of test collections. Section 6.2 describes the development

of a test collection for Tetun, while experiments with the use of LLMs for automated relevance judgments are presented in Section 6.3.

6.2 Constructing a Test Collection for Tetun

This section presents the construction of a test collection for Tetun, named *Labadain-Avaliadór* (*avaliadór*, a Tetun word meaning “evaluator”). It details the document collection process and the procedures for query formulation, document pooling, and relevance judgments.

6.2.1 Document Collection

The document collection comprises 33,550 Tetun documents sourced from the Labadain-30k+ dataset [64], each enriched with metadata including a title, URL, source, publication date, and content. This dataset was collected from web crawling and covers a wide range of categories, including news articles, Wikipedia entries, legal and government documents, research papers, technical documents, blogs, forums, and more [62]. The diversity of its sources and topics makes this dataset particularly suitable for constructing a Tetun test collection.

A sample of the documents, formatted according to TREC guidelines, is shown in Figure 6.1. The collection is 84 MB in size, with approximately 12.3 million tokens and 162,466 unique tokens. A summary of the collection is provided in Table 4.11, and the length distribution of titles and content is illustrated in Figure 6.2.

<pre> <doc> <docno>001-001-2022-4</docno> <title>Prizioneiru 400 resin ezerse direitu votu iha prizaun Bekora</title> <url>https://tatoli.tl/2022/03/19/prizioneiru-atus-hat-resin-ezerse-direitu-votu-ih-prizaun-becora/</url> <source>tatoli.tl</source> <category>Notísia</category> <published_date>2022-03-19</published_date> <text> DILI, 19 Marsu 2022 - Prizioneiru iha komarka Bekora hamutuk 474 ezerse direitu votu iha estabesimentu prizonál Bekora Dili. Prizioneiru hirak ne'ebé elejivel atu tuir votasaun hamutuk 474 maibé seidauk identifika loloos, tanba prizoneru hirak ne'ebé moras mentál karik bainhira sira kondisaun normál bele tuir no moras ne'ebé grave sei la posível atu tuir votasaun. Diretór estabesimentu prizonal, João Domingos, afirma sira sei la obriga prizioneiru hirak ne'ebé moras mentál. </text> </doc> </pre>	<pre> <doc> <docno>001-001-2022-4</docno> <title>Over 400 prisoners exercised their voting rights in Bekora prison</title> <url>https://tatoli.tl/2022/03/19/prizioneiru-atus-hat-resin-ezerse-direitu-votu-ih-prizaun-becora/</url> <source>tatoli.tl</source> <category>News</category> <published_date>2022-03-19</published_date> <text> DILI, 19 March 2022 - A total of 474 prisoners at Bekora prison exercised their voting rights within the Bekora prison facility in Dili. The eligible prisoners totalled 474, but the exact number has not yet been confirmed, as prisoners with mental health issues might be allowed to vote if they are in a stable condition, and those with severe conditions will not be able to participate in the voting. The director of the prison facility, João Domingos, stated that they will not force prisoners with mental illness. </text> </doc> </pre>
--	--

Figure 6.1: Sample of the document formatted following TREC guidelines: original (left) and English translation (right).

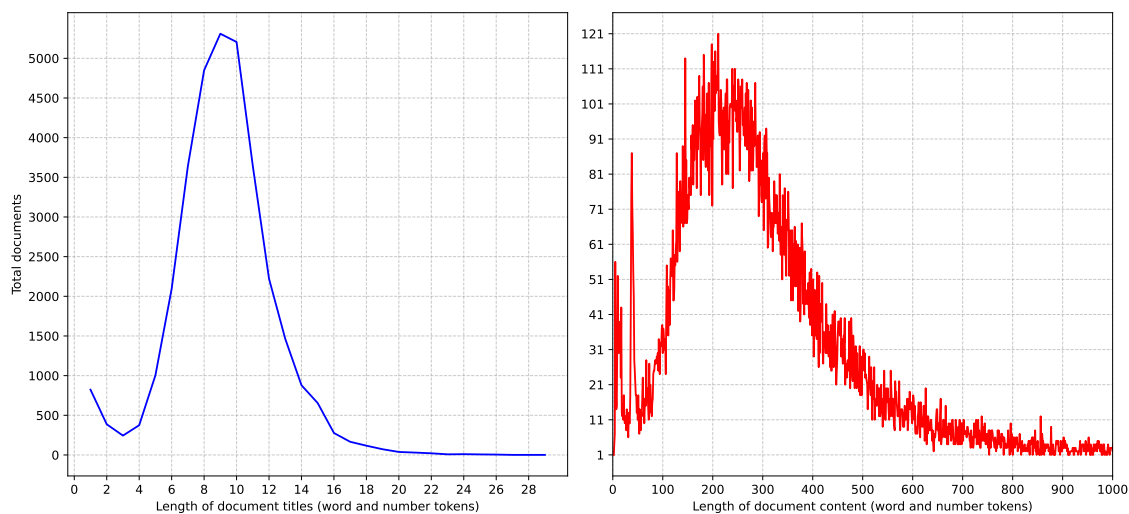


Figure 6.2: Length distribution of titles and body content in the Labadain-30k+ dataset. In the right-hand figure, the x-axis of the document length distribution is limited to 1,000 tokens (words and numbers) for improved visualization.

6.2.2 Query Formulation

Queries were collected from two sources: the Google Search Console¹ for Timor News and the user search logs from the Timor News platform [59]. The Google Search Console queries cover the period from 1 November 2021 to 31 January 2024, while the Timor News logs span from 7 May 2021 to 31 January 2024. The collected queries were compiled and distributed to five second-year undergraduate volunteers, all native Tetun speakers. The group consisted of two students from Environmental Engineering, two from Information Systems, and one from Medicine. These volunteers were tasked with developing queries following established guidelines.

Initially, each student was assigned 250 queries, resulting in a total of 1,250 queries analyzed. To understand user information needs, the students reviewed the provided search logs and either retained, modified, or formulated new queries based on the contextual information interpreted from the logs (see examples in Table 6.1). Before beginning query development, the students attended a training session that provided practical examples of query formulation. Following this, three pilot testing sessions were conducted, during which each student created a query, defined the associated information need, and described the types of documents they would consider relevant.

The queries were then entered into a search prototype built on top of Apache Solr using the BM25 ranking model. This prototype allowed the students to analyze the documents retrieved for each input query. For each query, each student selected four documents from the top 50 retrieved list, ensuring that one document represented each category: non-relevant, marginally relevant, relevant, or highly relevant. These sessions facilitated discussions about the results, highlighted challenges faced, and provided feedback to deepen their understanding of query development.

¹<https://search.google.com/search-console>

Table 6.1: Examples of original query logs and their reformulations. Words highlighted with a *green* background indicate newly added terms, while word with an *orange* background indicates orthography correction.

Original query log	Reformulated query
Problema lixu (Waste problem)	Problema lixu iha Dili (Waste problem in Dili)
Soe bebe (Baby abandonment)	Kazu soe bebé (Case of baby abandonment)
Konsumu tabaku (Tobacco consumption)	Dadus konsumu tabaku (Tobacco consumption data)
Kilat ilegal (Illegal firearms)	Problema kilat ilegal (Illegal firearms problem)

Subsequently, students were tasked to develop short queries according to TREC best practices [194], ranging from three to five words, specifying information needs and describing the types of documents they expected the system to retrieve. Using the search prototype mentioned earlier, they input their queries and analyzed the retrieved documents. To finalize a query, they ensured that at least five relevant documents could be identified. In total, 61 queries were developed, and a sample *topic*, formatted according to TREC guidelines, is shown in Figure 6.3. A summary of the queries is provided in Table 6.2, and their distribution across categories is shown in Figure 6.4. For categorization, we adapted the query topic categorization frameworks of Beitzel et al. [23] and Rohatgi et al. [190].

<pre> <top> <num> Number: TTI-00001 <title> Topic: Prevenσαun moras HIV-SIDA <desc> Description: Informasaun kona-ba mekanizmu prevenσαun moras HIV-SIDA, inklui moras HIV-SIDA ninia kauza. <narr> Narrative: Dokumentu relevante bainhira kontein informasaun kona-ba esforsu no mekanizmu sira hodi prevene HIV-SIDA. Informasaun kona-ba kauza hosi HIV-SIDA mós relevante. Informasaun kona-ba seksu ne'ebé la detalle kona-ba ninia ligasaun ho moras HIV-SIDA, la relevante. </top> </pre>	<pre> <top> <num> Number: TTI-00001 <title> Topic: Prevention of HIV-AIDS <desc> Description: Information about HIV-AIDS prevention mechanisms, including the causes of HIV-AIDS. <narr> Narrative: Relevant documents are those that contain information about efforts and mechanisms to prevent HIV-AIDS. Information about the causes of HIV-AIDS is also relevant. Information about sex that does not detail its connection to HIV-AIDS is not relevant. </top> </pre>
--	--

Figure 6.3: Sample of topic formatted following TREC guidelines: original (left) and English translation (right).

6.2.3 Document Pooling

Given that query-document relevance judgments are conducted by human assessors, it is not feasible to evaluate all documents in a large collection. To address this challenge, Spärck-Jones and

Table 6.2: Summary of queries.

Description	Value
Queries	61
Three-word queries	37
Four-word queries	22
Five-word queries	2
Average words per query	3.43

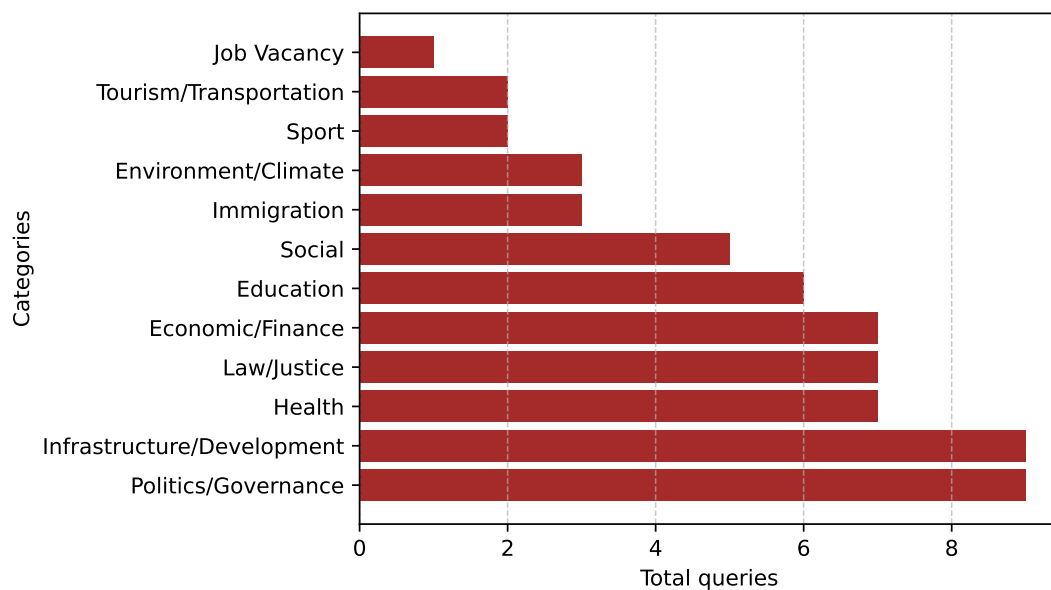


Figure 6.4: Distribution of queries across categories.

van Rijsbergen [210] introduced the *pooling* technique, in which a small subset containing a sufficiently representative sample of relevant documents is selected from the larger collection and provided to human assessors for relevance judgments [194].

For the development of the document pooling process, two probabilistic models were employed: BM25 and a language model (LM) with Dirichlet smoothing. BM25 is widely recognized for its strong performance across various IR tasks, particularly in ad-hoc retrieval scenarios. On the other hand, the language model with Dirichlet smoothing has demonstrated effectiveness in handling short queries [247]. By combining these models, the approach aims to maximize the retrieval of relevant documents and ensure the creation of a robust test collection.

Documents were indexed in separate instances of Solr, each configured with either the BM25 or the Dirichlet LM ranking models. The document retrieval and pooling process was fully automated and integrated with a relevance assessment interface to streamline the workflow. When a query was received, the system retrieved candidate documents ranked by each of the two models, applied balanced interleaving [183] to combine the results into a pool, and presented the top 100 documents to the assessors for relevance judgments. The architecture of the retrieval system used for this task is presented in Figure 6.5.

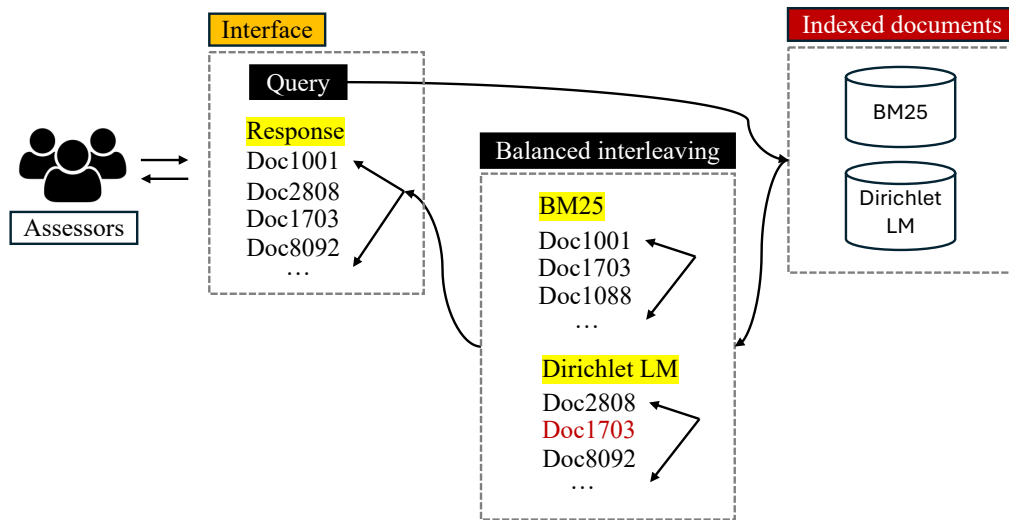


Figure 6.5: General architecture of the retrieval system used for relevance judgments. The document highlighted in *red* indicates a duplicate that was excluded from the final list after applying the balanced interleaving technique [183].

6.2.4 Relevance Judgments

The relevance judgment task was conducted by the same students who formulated the queries. We employed the four graded levels of topical relevance proposed by Sormunen [208]: non-relevant, marginally relevant, relevant, and highly relevant. Relevance judgments were performed following the guidelines outlined in Subsection 6.2.2. To streamline the process, a user-friendly web interface was developed, allowing assessors to log in with individual accounts to perform the judgments. The relevance judgment platform was made accessible to assessors through a secure public URL using Ngrok,² a tunneling service that facilitates the temporary, secure exposure of local web servers to the internet. The interface used for this task is shown in Figure 6.6.

Once logged in, the assessors initiated the judgments by selecting a query (label 1 in Figure 6.6) and reviewing the associated information need and relevance criteria (label 2 Figure 6.6). They then judged each of the 100 documents, assigning a relevance score to each, and submitted their judgments (label 3 Figure 6.6). After submission, the system redirected the assessors to the homepage, which displayed a list of queries. To maintain the integrity of the judgments, the option to reassess previously judged queries was automatically disabled.

Each assessor judged the same set of 61 queries, with 100 documents per query, resulting in a total of 6,100 documents being assessed. Assessors were instructed to focus solely on topical relevance, check document details if the query was too long and not fully displayed on the interface, and disregard the retrieval order when determining the document relevance for the given query. The assessment process was completed within eight hours, and the number of queries evaluated by each annotator per hour is illustrated in Figure 6.7. Any documents not included in the judgment list were automatically considered non-relevant.

²<https://ngrok.com>

Query-Documents Assessment Platform 1

This platform is designed to streamline the workflow of query-documents relevance Assessment.

List of Queries	
Sensu uma-kain 2022	Evaluate
Progresu dezentvolvimentu Infraestrutura	Evaluate
Preparasaun ba expo Dubai	Evaluate
Agresaun fizika hosi PNTL	Evaluate
Panorama eleisaun PR 2022	Evaluate
Inundasaun iha Dili	Evaluate

Query-Documents Relevance Assessment 2

Step 1 of 2: Query Description

Topic Sensu uma-kain 2022

Information need Informasaun kona-ba sensu uma-kain 2022.
Dokumentu sira ne'ebé fó sai informasaun kona-ba implementasaun no dadus estatistika ba sensu uma-kain 2022, relevante. Karik dokumentu kontein informasaun kona-ba sensu seluk, la relevante.

Relevant documents

Next

Query-Documents Relevance Assessment 3

Step 2 of 2: Evaluation and Submission

Uma-kain 40% partisipa ona sensu populasau 2022 iha Bobonaro

tatoli.tl
BOBONARO, 13 setembru 2022 — Diretór Estatistika munisípiu Bobonaro, Marti...

Relevance score: ☐ La relevante ☐ Relevante naton ☐ Relevante ☐ Relevante tebes

Uma-kain 21.6% iha territóriu partisipa ona sensu populasau 2022

tatoli.tl
DILI, 10 setembru 2022 —Governu liuhosi Diresaun Jerál Estatistika (DGE, s...

Relevance score: ☐ La relevante ☐ Relevante naton ☐ Relevante ☐ Relevante tebes

PM Taur no familia partisipa ona sensu populasau no uma-kain 20...

tatoli.tl
DILI, 30 setembru 2022 (TATOLI) —Primeiru-Ministru (PM), Taur Matan Ruak, ...

Relevance score: ☐ La relevante ☐ Relevante naton ☐ Relevante ☐ Relevante tebes

Abitante millaun 1 resin partisipa ona sensu uma-kain 2022

tatoli.tl
DILI, 07 outubru 2022 —Diresaun Jerál Estatistika (DGE, sigla portugues) ho...

Relevance score: ☐ La relevante ☐ Relevante naton ☐ Relevante ☐ Relevante tebes

Ministériu Finansas Lansa Rezultadu Prelimináriu Sensu Populasau...

timorpost.com
Dili – Governu Timor-Leste liuhusi Ministériu Finansas (MF), Rui Augusto G...

Figure 6.6: Web interface for human relevance assessments.

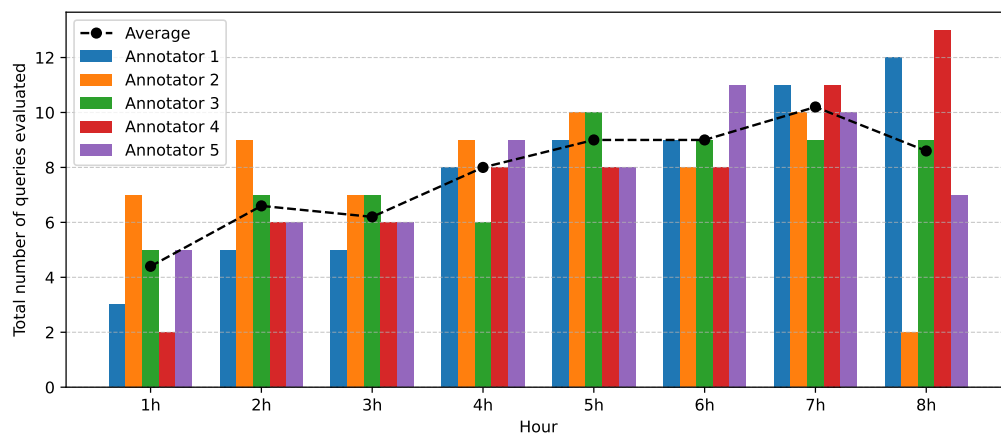


Figure 6.7: Hourly statistics of the total number of queries evaluated by each assessor.

To measure inter-annotator agreement, Cohen's kappa [43] was applied, with agreement strength interpreted according to the scale proposed by Landis and Koch [139]. The overall average score of inter-annotator agreement among the five annotators was 0.4236, indicating moderate agreement. Detailed results are provided in Table 6.3.

Since all assessors evaluated the same queries, the majority voting approach was used to determine each document's relevance to its corresponding query, based on the scores assigned to the query-document pairs. Under the majority voting rule, the most frequently chosen label for each query-document pair must exceed 50% of the total votes to qualify as the majority label [203], requiring at least three annotators to select the same label.

The evaluation results showed that 9.87% of the 6,100 judged documents did not meet this

Table 6.3: Cohen’s kappa scores for inter-annotator agreement among the five assessors.

	HA ₁	HA ₂	HA ₃	HA ₄	HA ₅
Human annotator 1 (HA ₁)	—	0.4344	0.4999	0.4434	0.4380
Human annotator 2 (HA ₂)	0.4344	—	0.3745	0.3310	0.3500
Human annotator 3 (HA ₃)	0.4999	0.3745	—	0.4646	0.4199
Human annotator 4 (HA ₄)	0.4434	0.3310	0.4646	—	0.4063
Human annotator 5 (HA ₅)	0.4380	0.3500	0.4199	0.4063	—
Average kappa score	0.4236				

criterion or resulted in ties. For example, ties occurred when two groups of annotators selected different labels, such as annotators 1 and 2 assigning score 3, annotators 3 and 4 assigning score 1, and the fifth annotator assigning score 2. Details of these tied documents are provided in Table 6.4.

Table 6.4: Summary of documents with tied scores.

Description	Value
Total number of tied documents	602
Proportion of tied documents	9.87%
Minimum number of tied documents per query	1
Maximum number of tied documents per query	38
Average number of tied documents per query	10.20
Standard deviation in the number of tied documents per query	7.17

To address tied scores, a tie-breaker strategy was applied, following the principle that using a strong signal to break ties is more effective than relying on a weak one [240]. Applying this method to all tied scores, the assessment results are presented in the “1st round column” of Table 6.5. However, after conducting an in-depth analysis of the tied scores, significant discrepancies were identified in certain cases, such as ties between scores of 0 and 2 or 1 and 3.

To resolve these inconsistencies, three of the five original assessors were re-invited for a second round of evaluations, focusing specifically on the tied documents. In this phase, assessors were presented with the two tied score options from the initial assessment. The reassessment process was conducted using Microsoft Excel, with separate tabs provided for each query and its corresponding documents. This second round was completed in approximately one hour and fifteen minutes.

After completing the second round of assessments, the evaluation results were merged with those from the first round, and the majority voting method was applied, selecting the most frequent score for each query-document pair as the final decision. To ensure reliability, queries with 100 relevant documents or fewer than ten relevant documents were excluded [8, 194], resulting in the removal of two queries with ten or fewer relevant documents in the second round.

The final test collection contains an average of 36.76 relevant documents per query, as detailed in Table 6.6. It comprises 9.59% highly relevant, 17.86% relevant, 9.31% marginally relevant, and

Table 6.5: Details of the human judgment results.

Relevance level	1st round		2nd round	
	#	%	#	%
3 - Highly relevant	710	11.64	566	9.59
2 - Relevant	1,102	18.07	1,054	17.86
1 - Marginally relevant	476	7.80	549	9.31
0 - Non-relevant	3,812	62.49	3,731	63.24

63.24% non-relevant documents, as shown in the 2nd round column in Table 6.5. The distribution of document relevance per query is illustrated in Figure 6.8.

Table 6.6: Summary of the final test collection. * Relevant documents consist of marginally relevant, relevant, and highly relevant.

Description	Value
Total number of topics	59
Total number of <i>qrels</i>	5,900
Minimum number of relevant documents per query*	11
Maximum number of relevant documents per query	99
Average number of relevant documents per query	36.76
The standard deviation of relevant documents per query	20.89

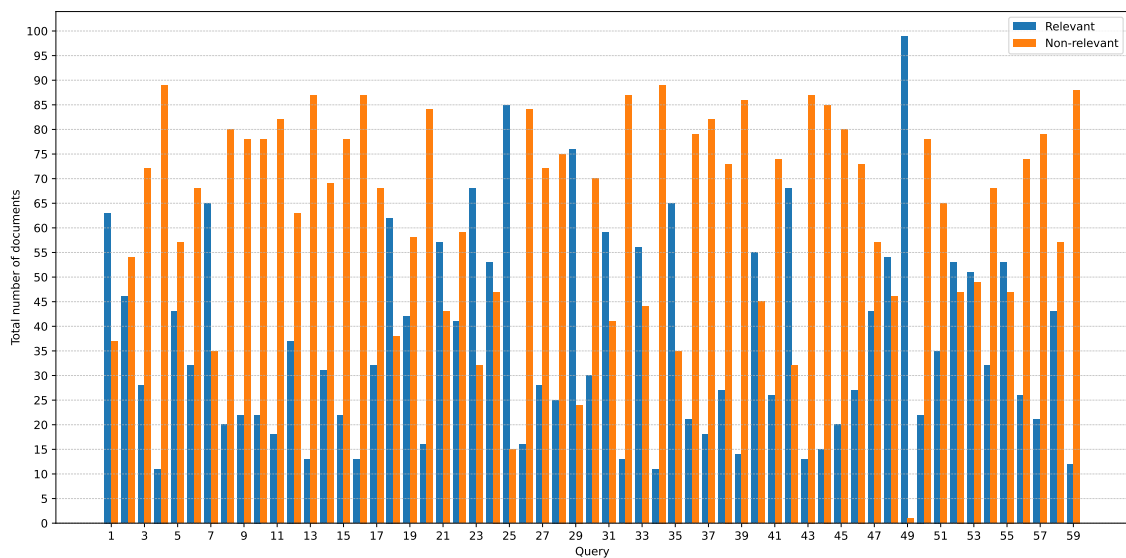


Figure 6.8: Total number of relevant and non-relevant documents per query. Relevant documents consist of marginally relevant, relevant, and highly relevant.

6.2.5 Results and Discussion

Table 6.5 shows that, after reassessing the tied documents from the first round results using the tie-breaker strategy, the number of highly relevant documents decreased by 144, and the relevant documents by 48. In contrast, marginally relevant documents increased by 73, while non-relevant documents decreased by 81. This shift suggests that some documents initially classified as highly relevant or relevant were labeled as marginally relevant, and some marginally relevant documents were reclassified as non-relevant. Furthermore, 200 documents from two excluded queries were eliminated, indicating a change in relevance interpretation after resolving the ties. These adjustments refined the distribution of documents across relevance categories, improving the overall robustness of the test collection.

When analyzing the changes in document relevance, *annotator 2* exhibited a distinct pattern, as illustrated in Figure 6.7. This annotator judged seven queries (700 documents) in the first hour at an average rate of 5.14 seconds per query-document pair, significantly faster than the average of approximately 8.18 seconds per pair observed among other annotators, who assessed an average of 4.4 queries (440 documents) in the first hour. This fast pace persisted until the seventh hour, leaving only two queries for the eighth hour.

Additionally, Table 6.3 shows that *annotator 2* had only fair agreement with three other annotators (3, 4, and 5), contributing to discrepancies. These observations indicate that *annotator 2* likely produced lower-quality annotations due to the fast-paced evaluation process, which affected the changes observed in document relevance after the second round of judgments.

Finally, Table 6.7 compares the *Labadain-Avaliadór* with the test collections of other LRLs. The *Labadain-Avaliadór* offers a balanced combination of scale and relevance density, making it a valuable resource for ad-hoc text retrieval in low-resource settings like Tetun. With 59 topics and 33,550 documents, it provides a moderately sized corpus that surpasses smaller collections like Sanskrit and Citumbuka but is smaller than larger collections like Hamshahri. Additionally, the diversity of its topics (see Figure 6.4) further ensures its suitability for ad-hoc retrieval tasks.

Table 6.7: Comparison of the *Labadain-Avaliadór* with other LRLs.

Language	#docs	#topics	Avg. relevant docs
Tetun	33,550	59	36.76
Sanskrit [191]	7,057	50	8.54
Chichewa [37]	9,380	129	19
Citumbuka [37]	2,258	129	17
Cinyanja [37]	173	129	15
Hamshahri [8]	166,774	65	36.18

The *Labadain-Avaliadór* stands out for its high relevance density, averaging 36.76 relevant documents per topic, the highest among the collections analyzed. This level of relevance provides a robust foundation for retrieval experiments, supported by a substantial pool of annotated relevance judgments that enable precise and reliable evaluations. Compared to collections with size

variations of up to 50,000 documents, such as Sanskrit and Chichewa, the *Labadain-Avaliadór* excels as a well-annotated collection, particularly suited for ad-hoc text retrieval in LRL contexts.

6.2.6 Conclusions

This study details the development of *Labadain-Avaliadór*, a Tetun test collection constructed following TREC guidelines. Five native Tetun-speaking students were responsible for both query development and query-document relevance judgments. The queries were derived from real-world Tetun search logs, and relevance judgments were graded on a scale from zero to four. The assessors reached agreement on more than 90% of the query-document pairs, with an average inter-annotator agreement of Cohen’s kappa score of 0.4236, indicating a moderate agreement.

Approximately 10% of the 6,100 query-document pairs resulted in disagreements, leading to tied scores. To resolve these ties, three of the five assessors conducted a second round of evaluation, selecting between two scoring options for each tied case. This process resulted in the creation of *Labadain-Avaliadór*, containing 5,900 *qrels*, categorized as 9.59% highly relevant, 17.86% relevant, 9.31% marginally relevant, and 63.24% non-relevant documents.

While the methodology for constructing test collections following TREC guidelines is well-established, there has been limited discussion in the literature on the specific challenges of creating test collections for LRLs. Key issues, such as resolving tied scores when developing graded collections in LRLs, remain insufficiently detailed. This study provides a detailed, step-by-step account of how a test collection can be developed for a new LRL. By addressing these gaps, it offers a practical and promising methodology for future IR researchers, serving as a framework for building test collections for other LRLs.

6.3 Exploring LLMs for Relevance Judgments in Tetun

To address the challenges of manually conducting relevance judgments, the IR community has explored the potential of automating this task. Studies have consistently shown that LLMs provide cost-effective solutions with judgment agreements comparable to human assessors. Faggioli et al. [89] argued that although further improvement in LLM capabilities is required for fully automated relevance judgments, LLMs are already capable of assisting humans in this task.

Additionally, a recent study by Bueno et al. [33] reported consistent improvements in automated relevance judgments with an average Cohen’s kappa score of 0.31 for annotation agreement between humans and LLMs, aligning with the findings of Faggioli et al. [89]. However, these studies primarily focus on high-resource languages, such as English and Brazilian Portuguese, leaving the applicability of LLMs in LRL contexts as an open question.

To address these challenges, this study explored an automated relevance judgment experiment for Tetun, using the test collection described in Section 6.2. The experiment employed a test collection containing 6,100 relevance judgments (corresponding to the first round in Table 6.5, as the second-round results were not yet finalized at that time). Query-document pairs were submitted to LLMs, tasked to assign relevance scores for each pair. These scores were then compared with

human annotations to calculate inter-annotator agreement levels. The details of the experiment are provided in the following subsections.

6.3.1 Overview

The fast pace of research and innovation, particularly with the emergence of LLMs, has significantly transformed NLP. Within the IR domain, studies have demonstrated that automated relevance judgments using LLMs can yield results comparable to traditional methods, and these outcomes have consistently improved as LLMs continue to evolve.

Initially, Faggioli et al. [89] explored the potential application of LLMs to fully automated relevance judgment tasks. They analyzed the judgment results from the TREC 2021 Deep Learning track [53] and compared them with LLM-based relevance assessments generated using GPT-3.5 of OpenAI.³ Their findings revealed a Cohen’s kappa score of 0.26 for inter-annotator agreement between humans and LLM, indicating a fair level of agreement. Thus, they argued that LLMs are already capable of assisting humans in relevance judgment tasks, despite further improvement in LLM capabilities being necessary for fully automating these tasks.

Later, Thomas et al. [217] reported that LLMs showed accuracy comparable to human labelers when deployed for large-scale relevance labeling at Bing.⁴ They used GPT-4 [172] and incorporated data from the TREC Robust04 track [230], showing that LLMs achieved a Cohen’s kappa score ranging from 0.20 to 0.64 for agreement between humans and LLM across various tasks.

More recently, Bueno et al. [33] in their study, while constructing a test collection for Brazilian Portuguese, reported consistent improvements and findings comparable to those of Thomas et al. [217] and Faggioli et al. [89], with automated relevance judgments yielding an average Cohen’s kappa score of 0.31 for annotation agreement between humans and LLMs. Similarly, this study reports an experiment conducted for Tetun in the context of LRLs.

6.3.2 Relevance Judgments for Tetun Using LLMs

Previous studies have investigated the use of OpenAI’s GPT-3.5 and GPT-4 models for automating relevance judgment tasks [33, 89, 217]. However, due to the costs associated with these LLMs, we explored an alternative solution by employing the freely available 70B variant of LLaMA3, released by Meta on April 18, 2024 [159].

To evaluate whether the free LLaMA3 model (70B variant) outperforms certain paid LLMs in relevance assessment tasks, particularly in Tetun, two paid models were selected for comparison: the Haiku variant of Claude 3 from Anthropic,⁵ and the Turbo variant of GPT-3.5 from OpenAI. A summary of the models and their associated costs is provided in Table 6.8.

To assess the suitability of the chosen LLMs for Tetun, including the two paid models, we conducted preliminary tests involving the translation of Tetun text into English. This step was

³<https://openai.com>

⁴<https://www.bing.com>

⁵<https://www.anthropic.com>

Table 6.8: Price associated with LLMs as of May 2024.

Model	Input	Output
LLaMA3 70B	—	—
Claude 3 Haiku	\$0.25 / 1M tokens	\$1.25 / 1M tokens
GPT-3.5-turbo-0125	\$0.50 / 1M tokens	\$1.50 / 1M tokens

essential, given that the query-document pairs are written in Tetun. Examples of these translated outputs are presented in Table 6.9, where inaccuracies in LLaMA3’s translation are highlighted with strike-through markings on two words.

Table 6.9: Examples of Tetun to English translations provided by LLMs.

Input Tetun text	Juventude hetan virus infiksaun HIV/SIDA tanba menus informasaun.
Model	Translation
LLaMA3 70B	Youth ~~against~~ virus infection of HIV/AIDS ~~and~~ lack of information.
Claude 3 Haiku	Youth get HIV/AIDS virus infection due to lack of information.
GPT-3.5-turbo-0125	The youth are getting infected with HIV/AIDS because of a lack of information.

To evaluate the quality of the translated documents generated by the LLMs, we randomly selected a sample of five documents from the query-document pairs (see examples in Table 6.10) and translated them into English by a native Tetun-speaking Ph.D. student. These human translations served as reference points for evaluation, with a single reference provided for each document. A statistical summary of the document lengths (in words) for the five randomly selected documents and their corresponding human translations into English is presented in Table 6.11.

Table 6.10: Examples of query-document pairs.

Query	Document
Prevensaun moras HIV/SIDA	UNFPA Sei Kooperera ho MS Hodi halo Prevensaun ba Moras HIV/SIDA
Prevensaun moras HIV/SIDA	KNK-HIV/SIDA Sensibiliza Informasaun HIV/SIDA Ba Trabalhador KSTL
Prevensaun moras HIV/SIDA	Autoridade Lokál Partisipa Workshop Prevensaun Moras HIV/SIDA

The five selected Tetun documents were provided to the LLMs to generate English translations. Both the machine-translated outputs and the reference translations (human-produced English translations) were preprocessed by lowercasing and then tokenized using the Tetun Simple

Table 6.11: Summary of document lengths (in words).

Language	Min	Max	Avg	Std
Tetun	8	13	10.0	1.79
English (translation)	11	15	12.6	1.36

Tokenizer [67] to create candidate and reference sets. These sets were subsequently evaluated using BLEU [175], demonstrated that both paid models outperformed LLaMA3 in translating Tetun to English, as shown in Table 6.12.

Table 6.12: Evaluation of translation quality using BLEU.

Model	BLEU
LLaMA3 70B	0.3849
Claude 3 Haiku	0.4167
GPT-3.5 Turbo	0.6525

However, given that relevance judgment tasks require not only direct translation but also a nuanced level of understanding, we compared the selected models’ multi-task language understanding capabilities using the Massive Multitask Language Understanding (MMLU) [110] based on the MMLU benchmark leaderboard provided by Papers with Code [237]. A summary of these LLMs’ performance on MMLU is shown in Table 6.13. The results indicate that in the few-shot scenario with five examples, LLaMA 3 outperformed Claude 3 Haiku by an average of +5 percentage points and GPT-3.5 Turbo by +10.2 percentage points.

Table 6.13: LLM capabilities of Multi-task Language Understanding on MMLU.

Model	Average (%)
LLaMA3 70B (5-shot)	80.2
Claude 3 Haiku (5-shot)	75.2
GPT-3.5 Turbo (5-shot)	70.0

6.3.3 Experiment with Tetun

To automate relevance judgments for Tetun using LLMs, we employed few-shot prompting, adapting a prompt structure similar to that proposed by Bueno et al. [33]. The prompt, along with an example, is illustrated in Prompt 6.3.1. A total of 6,100 query-document pairs were provided to the LLMs, which were tasked with assigning a relevance score to each pair. Examples of these query-document pairs are shown in Table 6.10.

Given that the existing Tetun test collection uses four-level relevance scores ranging from 0 to 3, we provided the LLMs with query-document pairs along with four examples, one for each relevance level. These examples were drawn from the same queries as those used in the pilot

testing phase by human assessors and included both the relevance score and the reasoning behind it. For each query-document pair, the LLMs were tasked with assigning one of the four scores and explaining the reasoning behind their choice.

Prompt 6.3.1: Example of the system prompt.

You are an expert assessor and you are tasked with assessing the relevance between the input query and its corresponding document, assigning a score from 0 to 3. A score of 0 indicates non-relevant; 1 is marginally relevant; 2 is relevant; and 3 is highly relevant.

Example:

query: “Kursu mestradu no pós-graduasaun UNTL”

document: “Kursu Desportu UNTL sei realiza graduasaun dahuluk tinan ne’e”

reason: “The query is about postgraduate and master’s courses at UNTL, whereas the document focuses on a sports course. Despite both courses in the query and document being offered at UNTL, the sports course in the document is not specifically designed for postgraduate or master’s levels. Thus, the document is only marginally relevant”.

score: 1

The query and document to be evaluated are the following:

query: {*query*}

document: {*document*}

Your response must be in JSON format with the first field is “reason”, explaining your reasoning, and the second field is “score”.

For the 70B variant of the LLaMA 3 model, which requires significant memory to run locally, a minimum of 40GB of RAM, as noted on Ollama,⁶ we leveraged the free API version of the cloud infrastructure provided by Groq⁷ to execute this model. However, the scripts for automated relevance judgments for all models were run locally.

The experiment began with the LLaMA3 70B model, as it was the primary focus for comparing annotator agreement levels with human assessors. We tested the model at temperatures of 0.0 and 0.5, following Ma et al. [152], who explored LLMs for relevance judgments in Chinese legal case retrieval. Increasing the temperature to 0.5 resulted in lower performance. Thus, a temperature setting of 0.0 was chosen for comparison with the other paid models.

⁶<https://ollama.com/library/llama3:70b>

⁷<https://console.groq.com/settings/billing>

6.3.4 Results and Discussion

In the experiment with the LLaMA3 70B model set at zero temperature, we obtained an inter-annotator agreement of Cohen’s kappa score of 0.2634 with human annotators. Increasing the temperature to 0.5 resulted in a slight decrease in inter-annotator agreement, with the kappa score dropping to 0.2594 (a reduction of -0.004). This observation aligns with findings by Ma et al. [152], where inter-annotator agreement levels between humans and LLMs also marginally decreased when the temperature was raised from 0.4 to 0.7 in evaluations of material facts.

Based on these results, we used a zero temperature setting when conducting relevance judgments using the Claude3 Haiku and GPT-3.5 Turbo models. Comparisons of the inter-annotator agreement levels between human annotators and LLMs are presented in Table 6.14.

Table 6.14: Cohen’s Kappa scores measuring the agreement between human assessors and LLMs.

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

The results show that the LLaMA3 70B model achieved the highest Cohen’s kappa score, indicating the strongest agreement with human annotators among the tested models. Among the paid models, GPT-3.5 Turbo exhibited a marginally higher kappa score than Claude3 Haiku, with an increase of +0.0012. Despite the superior performance of the paid models in translating Tetun into English, this finding suggests that deeper language understanding, rather than translation capability, is more crucial for automated relevance judgment tasks.

Our findings using the LLaMA3 70B model closely align with the initial results reported by Faggioli et al. [89] and are consistent with the findings of Bueno et al. [33] and Thomas et al. [217]. A comparison of these findings is presented in Table 6.15.

Table 6.15: Comparison of inter-annotator agreement levels in relevance judgment tasks.

	LLM	Cohen’s k Score
Bueno et al. [33]	GPT-4	0.31
Thomas et al. [217]	GPT-4	0.26 – 0.64
Faggioli et al. [89]	GPT 3.5	0.26
Ours	LLaMA3 70B	0.26

On average, our experiments required 3.56 hours to complete the relevance judgment tasks for each model. The costs associated with the two paid models are detailed in Table 6.16. Notably, GPT-3.5 Turbo was \$0.25 more expensive per one million input and output tokens compared to Claude3 Haiku, resulting in higher overall expenses for GPT-3.5 Turbo.

6.3.5 Conclusions

This study explored the use of LLMs for automating relevance judgment tasks in LRL scenarios, using Tetun as a case study. The results were comparable to those achieved in high-resource lan-

Table 6.16: Expense on LLMs for relevance judgment tasks.

Model	Expense
LLaMA3 70B	—
Claude 3 Haiku	\$1.98
GPT-3.5 Turbo	\$2.73

guages, underscoring the potential of LLMs for advancing retrieval research in LRL contexts. The availability of freely accessible models such as LLaMA3 further strengthens these opportunities.

Despite LLaMA3’s knowledge cutoff in December 2023 [5] and the presence of fewer than 45,000 Tetun documents on the web at that time [62, 136], the model achieved agreement levels comparable to those observed in English. This demonstrates the feasibility of automated relevance judgments for other LRLs as well.

As the first study to investigate automated relevance judgments with LLMs in LRL settings, this work provides a foundational contribution. It not only establishes a practical framework for future research but also encourages broader exploration of LLM-driven evaluation methods in under-resourced languages.

Chapter 7

Tetun Ad-Hoc Text Retrieval

This chapter presents the results of experiments on ad-hoc text retrieval in Tetun, drawing from three original papers produced during this research: an initial exploratory study conducted in the early stages of the doctoral journey [61], a study establishing a foundational baseline for Tetun ad-hoc text retrieval [73], and an investigation of pretrained dense retrieval models applied to Tetun ad-hoc text retrieval in zero-shot and hybrid settings [77].

The chapter provides a comprehensive overview of the experimental pipeline, examining the impact of text preprocessing and identifying the strategies that yielded effective retrieval solutions. It begins with preliminary findings, proceeds to baseline results, and concludes with insights from adapting pretrained dense retrieval models to Tetun under zero-shot and hybrid strategies.

7.1 Preliminary Study

In the early stages of this research, the resources for Tetun described in Chapters 5 and 6 were not yet available. As a result, a limited document collection and existing tools were used to obtain preliminary results for preprocessing, indexing, and retrieval. This initial study laid the groundwork for subsequent research in Tetun ad-hoc text retrieval.

7.1.1 Overview

Since foundational Tetun resources, such as the Labadain-30k+ dataset [64], the Tetun stopword list [72], and the Tetun stemmer [75], had not yet been developed, this study relied on the limited resources available at the time. It used 442 documents and five queries sourced from Timor News [59]. The document collection is news articles containing a title, a lead (a summary paragraph), and the main content, while the queries are drawn from the platform’s search logs. Apache Solr was employed for indexing, with BM25 (configured with its default settings) as the retrieval and ranking model.

To enable text processing, a list of 70 Tetun stopwords was manually created by translating English stopwords from the NLTK library into Tetun. A simple rule-based stemmer was implemented to handle noun forms ending in “*saun*”, “*mentu*”, “*dór*”, and “*-teen*” to reduce them to

their root forms. Text normalization rules were also applied to correct misspelled Tetun terms, such as [“junho”, “junu”] to “*juñu*” (june). These rules were further extended to generate abbreviations (e.g., *pm* to *primeiru ministru*) and synonym expansion lists based on information extracted from Timor News query logs.

7.1.2 Experimental Setup

The experiments were conducted using three configurations for query and document processing: default, with stopwords, and without stemming. The *default* configuration applied preprocessing techniques including HTML tag removal, lowercasing, stopword removal, abbreviation and synonym expansion, and stemming. The *with stopwords* configuration followed the default setup, except that stopword removal was not applied (i.e., stopwords were retained). The *without stemming* configuration also followed the default setup, but without applying stemming.

Four document element combinations for indexing were tested and evaluated: (i) title, (ii) title and content, (iii) lead and content, and (iv) title, lead, and content. Combining the three document processing approaches with the four indexing strategies resulted in a total of 12 distinct retrieval strategies for the experiments (see [Table 7.1](#)).

7.1.3 Test Collection Development

The test collection for evaluation was developed by pooling the top ten ranked documents from the results of 12 retrieval strategies for each query using the balanced interleaving approach [183], resulting in a total of 50 query-document pairs. These pairs were provided to human assessors for relevance judgments, following the established guidelines.

A four-level relevance scoring scale, as reported in the TREC 2020 News Track [207], was employed, with scores ranging from 0 to 3, defined as follows:

- 0 - The document provides *no useful information* (not relevant).
- 1 - The document provides *some useful information* or contextual details that help the user understand the broader context of the query (less relevant).
- 2 - The document provides *significant information* (relevant).
- 3 - The document provides *essential information* (highly relevant).

The 50 query-document pairs were assigned to five independent native Tetun-speaking assessors. Each query was associated with ten associated documents, and the assessors evaluated the query-document relevance using the specified scoring scale. The relevance judgments from the assessors were compiled using majority voting to determine a final score for each query-document pair. The resulting test collection included 442 documents, 5 topics, and 50 *qrels*.

7.1.4 Evaluation and Results

In the evaluation, the *results* from 12 retrieval strategies and *qrels* were compiled and formatted according to TREC format. The `trec_eval`¹ tool was used to evaluate retrieval effectiveness.

Table 7.1: The evaluation results.

Configuration	Document element	P@5	P@10	MAP	NDCG
Default	title	0.800	0.533	0.709	0.775
	title + content	0.667	0.567	0.662	0.715
	lead + content	0.633	0.533	0.649	0.780
	title + lead + content	0.633	0.517	0.652	0.732
With stopwords	title	0.600	0.483	0.572	0.715
	title + content	0.600	0.533	0.612	0.705
	lead + content	0.600	0.500	0.593	0.688
	title + lead + content	0.667	0.517	0.648	0.729
Without stemming	title	0.800	0.550	0.731	0.790
	title + content	0.700	0.550	0.671	0.730
	lead + content	0.700	0.517	0.650	0.719
	title + lead + content	0.700	0.550	0.650	0.738

The evaluation metrics included precision at the top 5 (P@5) and 10 (P@10), MAP, and NDCG. The results showed that indexing news titles without stemming document and query terms yielded the best performance, achieving a 3.1% relative improvement in overall MAP and 1.9% in overall NDCG compared to performance with stemming (*default* setting), as shown in Table 7.1.

7.1.5 Conclusions

This study reports preliminary findings from an early ad-hoc text retrieval experiment using Tetun documents and queries sourced from Timor News. Several indexing and retrieval strategies were evaluated to assess their effectiveness for Tetun text retrieval.

The results demonstrate that title-only retrieval without stemming is the most effective strategy, underscoring the crucial role of document titles in retrieval effectiveness. Stopword removal proved beneficial, while incorporating document leads and full content did not improve performance. Stemming also appeared unnecessary, as retrieval scores remained strong without it. The superior performance of title-only retrieval highlights the importance of query terms appearing in page titles, consistent with findings that such occurrences correlate with higher user engagement and click-through rates [17].

¹https://trec.nist.gov/trec_eval/

These findings provide an initial foundation for Tetun text retrieval research by identifying key factors that influence retrieval performance. It provides insights into retrieval behavior in Tetun and serves as a basis for refining indexing and retrieval strategies in subsequent works.

7.2 Baselines

Building on the findings of the preliminary study and the resources developed in Chapters 5 and 6, this section establishes baselines for ad-hoc text retrieval in Tetun using classical retrieval and ranking models. The goal is to evaluate the impact of different preprocessing, indexing, and retrieval strategies on retrieval effectiveness. Performance is measured using key metrics such as precision, MAP, and NDCG, providing a foundation for future research and advancements in Tetun ad-hoc text retrieval.

7.2.1 Overview

Research in ad-hoc text retrieval has demonstrated the effectiveness of various retrieval and ranking models. BM25, a widely recognized probabilistic model, is well established as a strong baseline in classical IR systems [188]. Likewise, the probabilistic language model (LM) with Dirichlet smoothing [155] has demonstrated robust performance, particularly for short queries in ad-hoc text retrieval tasks [247].

The divergence from randomness (DFR) variant of BM25 (DFR BM25) has shown competitive performance in various retrieval settings and has been validated in multiple TREC experiments [176, 238]. Similarly, the Hiemstra LM [111, 112] has performed equally well in ad-hoc text retrieval, especially in LRL scenarios [191]. These retrieval and ranking models form the basis of the experiments conducted in this study.

7.2.2 Text Preprocessing

Given the language-specific characteristics of Tetun, the text preprocessing workflow was organized into the following stages:

1. **Standard preprocessing:** This stage involved basic text cleaning operations, including lowercasing, normalizing apostrophes, removing punctuation and special characters, performing tokenization, removing extra spaces, and filtering out words longer than 60 characters.
2. **Language-specific preprocessing:** To address Tetun's unique linguistic features, additional text preprocessing techniques were applied, such as removing apostrophes, accents, and hyphens to split hyphen-connected words. Each of these steps was implemented independently within the preprocessing workflow.
3. **Stopword Removal and Stemming:** This stage extended beyond character-level preprocessing by incorporating stopwords removal and stemming. The *Labadain-Stopwords* list

was employed for stopwords removal, and the *Labadain-Stemmer* with three variants (light, moderate, and heavy) was applied for stemming.

7.2.3 Experimental Setup

Several retrieval strategies were explored by applying multiple text preprocessing techniques to assess their impact on retrieval effectiveness. The experiment workflow is illustrated in Figure 7.1.

The baseline was first established using the standard preprocessing step, as described in Subsection 7.2.2. Following this, individual language-specific text preprocessing techniques, including stopwords removal and stemming, were tested to compare their effectiveness. Finally, techniques outperforming the baseline were combined and further evaluated to identify the most effective strategies for Tetun ad-hoc text retrieval.

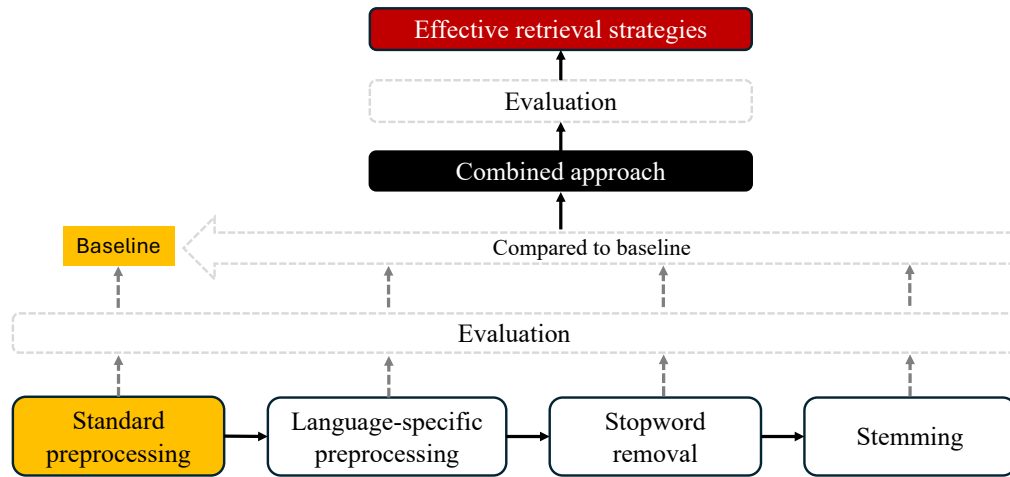


Figure 7.1: Overview of the experimental workflow.

For stemming, although the *moderate* and *heavy* variants outperformed the *light* variant in intrinsic assessment, the difference was minimal. Moreover, Flores and Moreira [93] observed that the most accurate stemmer does not always lead to the most effective retrieval. Therefore, all stemmer variants were included in the experiments to evaluate their impact.

We used PyTerrier [154], a Python API for the Terrier IR platform [173], for indexing, retrieval, and ranking, with the default settings maintained for each model. The same text preprocessing techniques were applied to both queries and documents, and the experiments were conducted using document titles and content to evaluate retrieval effectiveness for each strategy.

7.2.4 Document Indexing

Document titles and content were indexed independently, with the index created from standard preprocessing steps serving as the baseline. Additional indexes were generated for individual results obtained from applying language-specific preprocessing steps, stopwords removal, and each of the stemmer variants. To evaluate the effects of these preprocessing methods on index size,

index compression factors (ICF) [95] were calculated for each preprocessing configuration. A summary of the index compression results is presented in Table 7.2.

Table 7.2: Index compression factor (%) for titles and content compared to the baseline.

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

The results demonstrate that all preprocessing methods reduced the index size compared to the baseline. Removing hyphens achieved the highest compression factor for title indexing, reducing the index size by up to 30.76% relative to the baseline. For content indexing, the *heavy* stemmer variant resulted in the most significant compression, reducing the index size by up to 12.18%.

As expected, the *moderate* stemmer variant compressed the index more efficiently than the *light* variant, while the *heavy* variant achieved the highest compression for both title and content indexing. Interestingly, removing apostrophes and accents also contributed to index size reduction, with accent removal achieving over a 7% reduction for both titles and content. This suggests a substantial presence of identical words with and without accents in the documents.

7.2.5 Short-Text Retrieval Results

For short-text retrieval, document titles were indexed and used for retrieval. The impact of each preprocessing technique on retrieval effectiveness is presented Table 7.3. Scores highlighted in red indicate values lower than the baseline. As observed, removing accents and applying all stemmer variants did not improve retrieval effectiveness compared to the baseline.

Stopword removal yielded inconsistent results across models and metrics. It generally demonstrated improved performance only at the top-20 cutoffs (P@20, MAP@20, and NDCG@20) and overall MAP and NDCG, while exhibiting lower performance at top-5 and top-10 cutoffs across all metrics. Regarding the impact of stemming on retrieval effectiveness, the *light* stemmer variant slightly outperformed the *moderate* and *heavy* variants across all cutoffs, with the *moderate* and *heavy* variants performing similarly.

Based on the preliminary results in Table 7.3, all combinations of preprocessing techniques that outperformed the baseline, including stopword removal and the *light* stemmer variant, were selected for further comparison. The best results from these combined preprocessing techniques are presented in Table 7.4. The findings indicate that removing apostrophes and hyphens significantly enhances retrieval performance compared to the baseline.

Among the retrieval models, DFR BM25 consistently delivered the highest performance across all metrics up to top-10 cutoffs. Notable scores include P@5 (0.8881), P@10 (0.8390), MAP@5 (0.1589, removing stopwords), MAP@10 (0.2804), NDCG@5 (0.7512), and NDCG@10 (0.7356). DFR BM25 demonstrated relative improvements of up to 30.35% over the baseline and up to 8.19% over individual preprocessing techniques on multiple metrics, along with modest relative gains of up to 5.54% in MAP@5 over other retrieval models within the same settings. While DFR

Table 7.3: Effectiveness of text preprocessing techniques in short-text retrieval. Red values indicate scores lower than the baseline.

Retrieval Strategies	Model	Precision at Cutoff			MAP at Cutoff			NDCG at Cutoff			MAP	NDCG
		@5	@10	@20	@5	@10	@20	@5	@10	@20		
Baseline	BM25	0.8169	0.7763	0.6602	0.1444	0.2568	0.3903	0.6801	0.6668	0.6454	0.5925	0.7408
	DFR BM25	0.8169	0.7763	0.6619	0.1440	0.2563	0.3901	0.6811	0.6666	0.6468	0.5926	0.7407
	TF-IDF	0.8136	0.7746	0.6458	0.1432	0.2546	0.3825	0.6739	0.6640	0.6380	0.5802	0.7364
	Dirichlet LM	0.7898	0.7525	0.6398	0.1299	0.2361	0.3671	0.6359	0.6356	0.6174	0.5780	0.7208
	Hiemstra LM	0.8136	0.7695	0.6669	0.1428	0.2521	0.3928	0.6670	0.6588	0.6465	0.6090	0.7435
Remove apostrophes	BM25	0.8237	0.7763	0.6644	0.1453	0.2572	0.3930	0.6866	0.6685	0.6499	0.5938	0.7419
	DFR BM25	0.8237	0.7763	0.6661	0.1450	0.2568	0.3929	0.6878	0.6684	0.6515	0.5942	0.7420
	TF-IDF	0.8203	0.7746	0.6500	0.1443	0.2552	0.3854	0.6808	0.6660	0.6428	0.5818	0.7377
	Dirichlet LM	0.7898	0.7542	0.6432	0.1301	0.2365	0.3686	0.6380	0.6376	0.6206	0.5794	0.7219
	Hiemstra LM	0.8169	0.7712	0.6712	0.1429	0.2529	0.3953	0.6725	0.6609	0.6507	0.6102	0.7443
Remove hyphens	BM25	0.8271	0.7881	0.6856	0.1459	0.2616	0.4069	0.7143	0.7014	0.6871	0.6498	0.8130
	DFR BM25	0.8271	0.7881	0.6856	0.1459	0.2616	0.4070	0.7138	0.7016	0.6873	0.6506	0.8135
	TF-IDF	0.8271	0.7814	0.6805	0.1457	0.2573	0.4028	0.7118	0.6979	0.6845	0.6402	0.8077
	Dirichlet LM	0.7898	0.7576	0.6797	0.1322	0.2420	0.3860	0.6578	0.6615	0.6652	0.6679	0.8039
	Hiemstra LM	0.8339	0.7881	0.6898	0.1472	0.2635	0.4143	0.7142	0.6980	0.6914	0.6841	0.8239
Remove accents	BM25	0.7831	0.7542	0.6424	0.1370	0.2445	0.3735	0.6564	0.6512	0.6308	0.5744	0.7329
	DFR BM25	0.7831	0.7542	0.6441	0.1366	0.2440	0.3734	0.6566	0.6509	0.6317	0.5747	0.7328
	TF-IDF	0.7864	0.7508	0.6297	0.1364	0.2417	0.3666	0.6555	0.6477	0.6237	0.5631	0.7287
	Dirichlet LM	0.7390	0.7237	0.6331	0.1159	0.2173	0.3503	0.5945	0.6081	0.6036	0.5588	0.7104
	Hiemstra LM	0.7763	0.7441	0.6458	0.1341	0.2388	0.3732	0.6446	0.6406	0.6305	0.5880	0.7352
Remove stopwords	BM25	0.8102	0.7729	0.6695	0.1438	0.2547	0.3976	0.6693	0.6600	0.6488	0.6030	0.7443
	DFR BM25	0.8102	0.7729	0.6712	0.1439	0.2549	0.3984	0.6695	0.6602	0.6503	0.6049	0.7451
	TF-IDF	0.8102	0.7729	0.6686	0.1439	0.2549	0.3975	0.6691	0.6600	0.6484	0.6018	0.7438
	Dirichlet LM	0.8034	0.7593	0.6653	0.1317	0.2379	0.3803	0.6329	0.6315	0.6299	0.5936	0.7255
	Hiemstra LM	0.8203	0.7678	0.6864	0.1437	0.2521	0.4036	0.6702	0.6587	0.6577	0.6189	0.7483
Light stemming	BM25	0.8000	0.7678	0.6500	0.1381	0.2464	0.3758	0.6693	0.6605	0.6355	0.5826	0.7364
	DFR BM25	0.8000	0.7661	0.6492	0.1376	0.2456	0.3748	0.6679	0.6588	0.6345	0.5824	0.7358
	TF-IDF	0.7966	0.7610	0.6407	0.1371	0.2421	0.3700	0.6648	0.6551	0.6304	0.5711	0.7331
	Dirichlet LM	0.7661	0.7203	0.6305	0.1225	0.2198	0.3518	0.6286	0.6192	0.6076	0.5613	0.7174
	Hiemstra LM	0.7898	0.7492	0.6576	0.1315	0.2365	0.3760	0.6456	0.6410	0.6326	0.5907	0.7344
Moderate stemming	BM25	0.7797	0.7610	0.6466	0.1336	0.2423	0.3714	0.6594	0.6553	0.6319	0.5790	0.7346
	DFR BM25	0.7797	0.7593	0.6458	0.1331	0.2415	0.3704	0.6581	0.6535	0.6309	0.5789	0.7341
	TF-IDF	0.7763	0.7542	0.6373	0.1326	0.2380	0.3656	0.6550	0.6499	0.6268	0.5676	0.7313
	Dirichlet LM	0.7593	0.7186	0.6254	0.1197	0.2175	0.3484	0.6225	0.6149	0.6025	0.5577	0.7138
	Hiemstra LM	0.7729	0.7390	0.6525	0.1275	0.2319	0.3710	0.6365	0.6337	0.6277	0.5868	0.7322
Heavy stemming	BM25	0.7797	0.7610	0.6466	0.1336	0.2423	0.3714	0.6594	0.6553	0.6319	0.5788	0.7346
	DFR BM25	0.7797	0.7593	0.6458	0.1333	0.2416	0.3705	0.6583	0.6537	0.6311	0.5788	0.7341
	TF-IDF	0.7763	0.7542	0.6373	0.1326	0.2380	0.3656	0.6550	0.6499	0.6268	0.5674	0.7312
	Dirichlet LM	0.7559	0.7186	0.6254	0.1189	0.2170	0.3479	0.6200	0.6145	0.6022	0.5571	0.7135
	Hiemstra LM	0.7729	0.7390	0.6534	0.1275	0.2319	0.3715	0.6365	0.6337	0.6282	0.5867	0.7322

BM25 achieved the same P@5 as BM25 and MAP@5 as TF-IDF, it demonstrated slightly higher scores in other cutoffs.

The removal of hyphens and stopwords also proved beneficial, particularly with Hiemstra LM, which demonstrated the highest scores for P@20 (0.7305), MAP@20 (0.4372), NDCG@20 (0.7152), and overall MAP (0.7040) and NDCG (0.8289, with apostrophe removal). This combination enhanced retrieval effectiveness overall, with Hiemstra LM consistently showing relative improvements at top-20 cutoffs and in global MAP and NDCG. It demonstrated relative improvements of up to 15.60% over the baseline and up to 8.33% over individual text preprocessing techniques in multiple metrics, with marginal relative gains of up to 3.94% over other retrieval models within the same setting.

Stemming, however, had minimal impact on retrieval effectiveness, even when combined with other text preprocessing techniques that outperformed the baseline. The best result involving the *light* stemmer variant was achieved when paired with hyphen and stopword removal. While this

Table 7.4: Effectiveness of combined text preprocessing techniques in short-text retrieval. Values highlighted with a *green* background indicate the best score at the respective metric and cutoff.

Retrieval Strategies	Model	Precision at Cutoff			MAP at Cutoff			NDCG at Cutoff			MAP	NDCG
		@5	@10	@20	@5	@10	@20	@5	@10	@20		
Baseline	BM25	0.8169	0.7763	0.6602	0.1444	0.2568	0.3903	0.6801	0.6668	0.6454	0.5925	0.7408
	DFR BM25	0.8169	0.7763	0.6619	0.1440	0.2563	0.3901	0.6811	0.6666	0.6468	0.5926	0.7407
	TF-IDF	0.8136	0.7746	0.6458	0.1432	0.2546	0.3825	0.6739	0.6640	0.6380	0.5802	0.7364
	Dirichlet LM	0.7898	0.7525	0.6398	0.1299	0.2361	0.3671	0.6359	0.6356	0.6174	0.5780	0.7208
	Hiemstra LM	0.8136	0.7695	0.6669	0.1428	0.2521	0.3928	0.6670	0.6588	0.6465	0.6090	0.7435
Remove apostrophes	BM25	0.8237	0.7763	0.6644	0.1453	0.2572	0.3930	0.6866	0.6685	0.6499	0.5938	0.7419
	DFR BM25	0.8237	0.7763	0.6661	0.1450	0.2568	0.3929	0.6878	0.6684	0.6515	0.5942	0.7420
	TF-IDF	0.8203	0.7746	0.6500	0.1443	0.2552	0.3854	0.6808	0.6660	0.6428	0.5818	0.7377
	Dirichlet LM	0.7898	0.7542	0.6432	0.1301	0.2365	0.3686	0.6380	0.6376	0.6206	0.5794	0.7219
	Hiemstra LM	0.8169	0.7712	0.6712	0.1429	0.2529	0.3953	0.6725	0.6609	0.6507	0.6102	0.7443
Remove hyphens	BM25	0.8271	0.7881	0.6856	0.1459	0.2616	0.4069	0.7143	0.7014	0.6871	0.6498	0.8130
	DFR BM25	0.8271	0.7881	0.6856	0.1459	0.2616	0.4070	0.7138	0.7016	0.6873	0.6506	0.8135
	TF-IDF	0.8271	0.7814	0.6805	0.1457	0.2573	0.4028	0.7118	0.6979	0.6845	0.6402	0.8077
	Dirichlet LM	0.7898	0.7576	0.6797	0.1322	0.2420	0.3860	0.6578	0.6615	0.6652	0.6679	0.8039
	Hiemstra LM	0.8339	0.7881	0.6898	0.1472	0.2635	0.4143	0.7142	0.6980	0.6914	0.6841	0.8239
Remove stopwords	BM25	0.8102	0.7729	0.6695	0.1438	0.2547	0.3976	0.6693	0.6600	0.6488	0.6030	0.7443
	DFR BM25	0.8102	0.7729	0.6712	0.1439	0.2549	0.3984	0.6695	0.6602	0.6503	0.6049	0.7451
	TF-IDF	0.8102	0.7729	0.6686	0.1439	0.2549	0.3975	0.6691	0.6600	0.6484	0.6018	0.7438
	Dirichlet LM	0.8034	0.7593	0.6653	0.1317	0.2379	0.3803	0.6329	0.6315	0.6299	0.5936	0.7255
	Hiemstra LM	0.8203	0.7678	0.6864	0.1437	0.2521	0.4036	0.6702	0.6587	0.6577	0.6189	0.7483
Remove apostrophes and hyphens	BM25	0.8881	0.8373	0.7153	0.1553	0.2796	0.4304	0.7500	0.7347	0.7133	0.6648	0.8213
	DFR BM25	0.8881	0.8390	0.7169	0.1553	0.2804	0.4313	0.7512	0.7356	0.7149	0.6664	0.8219
	TF-IDF	0.8780	0.8322	0.7119	0.1543	0.2759	0.4273	0.7401	0.7288	0.7086	0.6553	0.8149
	Dirichlet LM	0.8407	0.8034	0.7068	0.1390	0.2561	0.4099	0.6834	0.6920	0.6829	0.6713	0.8018
	Hiemstra LM	0.8780	0.8305	0.7263	0.1524	0.2743	0.4339	0.7379	0.7245	0.7147	0.6955	0.8282
Remove hyphens and stopwords	BM25	0.8814	0.8237	0.7237	0.1576	0.2720	0.4356	0.7394	0.7221	0.7130	0.6752	0.8220
	DFR BM25	0.8847	0.8254	0.7237	0.1585	0.2729	0.4366	0.7416	0.7228	0.7139	0.6764	0.8224
	TF-IDF	0.8847	0.8237	0.7229	0.1585	0.2722	0.4355	0.7416	0.7220	0.7126	0.6715	0.8202
	Dirichlet LM	0.8508	0.8102	0.7220	0.1409	0.2549	0.4171	0.6933	0.6925	0.6918	0.6863	0.8065
	Hiemstra LM	0.8746	0.8305	0.7305	0.1524	0.2720	0.4372	0.7294	0.7211	0.7152	0.7040	0.8288
Remove hyphens, apostrophes, and stopwords	BM25	0.8814	0.8220	0.7212	0.1580	0.2725	0.4342	0.7395	0.7211	0.7123	0.6743	0.8223
	DFR BM25	0.8847	0.8237	0.7212	0.1589	0.2735	0.4353	0.7418	0.7218	0.7131	0.6756	0.8228
	TF-IDF	0.8847	0.8220	0.7203	0.1589	0.2727	0.4341	0.7417	0.7210	0.7118	0.6705	0.8205
	Dirichlet LM	0.8508	0.8068	0.7144	0.1418	0.2546	0.4142	0.6949	0.6913	0.6872	0.6831	0.8053
	Hiemstra LM	0.8746	0.8254	0.7263	0.1528	0.2711	0.4353	0.7287	0.7190	0.7134	0.7029	0.8289
Remove hyphens and stopwords, and apply light stemmer	BM25	0.8678	0.8169	0.7110	0.1540	0.2698	0.4267	0.7367	0.7203	0.7053	0.6602	0.8154
	DFR BM25	0.8712	0.8153	0.7110	0.1549	0.2695	0.4268	0.7389	0.7200	0.7058	0.6608	0.8157
	TF-IDF	0.8712	0.8169	0.7102	0.1549	0.2701	0.4267	0.7389	0.7201	0.7048	0.6584	0.8145
	Dirichlet LM	0.8339	0.7932	0.7034	0.1390	0.2503	0.4056	0.6936	0.6860	0.6800	0.6622	0.8023
	Hiemstra LM	0.8576	0.8203	0.7229	0.1496	0.2678	0.4289	0.7256	0.7171	0.7083	0.6831	0.8208
Average relative performance gains of the best model compared to the baseline		8.72%	8.08%	9.54%	10.66%	9.40%	11.30%	10.29%	30.35%	10.63%	15.60%	11.49%
Average relative performance gains of the best model compared to individual text preprocessing techniques		7.60%	7.39%	6.16%	9.75%	8.19%	6.93%	7.23%	7.45%	6.09%	8.33%	7.58%
Average relative performance gains of the best model compared to others within the same retrieval strategy		2.65%	1.62%	1.03%	5.54%	3.41%	1.43%	3.35%	2.22%	1.06%	3.94%	1.38%

combination produced competitive results, it did not surpass the performance of strategies that excluded stemming.

7.2.6 Long-Text Retrieval Results

Document content was indexed and used for long-text retrieval. The impact of each preprocessing technique on retrieval effectiveness is summarized in Table 7.5. The results differ from those of short-text retrieval: apostrophe and accent removal did not provide any measurable benefit for

BM25-based models in terms of Precision or MAP relative to the baseline. Likewise, accent removal offered no improvement in P@10, P@20, or MAP when using the Dirichlet LM model.

However, there was evidence of improvement with Hiemstra LM in MAP and NDCG at various cutoffs and in overall NDCG. In contrast, hyphen removal, stopword removal, and stemming demonstrated clear advantages for long-text retrieval. Among the stemming variants, the *light* stemmer consistently outperformed the *moderate* and *heavy* variants, though the margin of improvement over the *moderate* variant was minimal.

Table 7.5: Effectiveness of text preprocessing techniques in long-text retrieval. Red values indicate scores lower than the baseline.

Retrieval Strategies	Model	Precision at Cutoff			MAP at Cutoff			NDCG at Cutoff			MAP	NDCG
		@5	@10	@20	@5	@10	@20	@5	@10	@20		
Baseline	BM25	0.4847	0.4525	0.3839	0.0769	0.1281	0.1931	0.3883	0.3800	0.3765	0.3429	0.5764
	DFR BM25	0.4746	0.4475	0.3839	0.0758	0.1259	0.1925	0.3826	0.3763	0.3753	0.3416	0.5754
	TF-IDF	0.5288	0.4746	0.4110	0.0855	0.1382	0.2068	0.4275	0.4086	0.4045	0.3564	0.5927
	Dirichlet LM	0.4576	0.4186	0.3669	0.0696	0.1111	0.1701	0.3577	0.3544	0.3545	0.3110	0.5558
	Hiemstra LM	0.5390	0.4915	0.4314	0.0856	0.1402	0.2138	0.4289	0.4158	0.4189	0.3655	0.5990
Remove apostrophes	BM25	0.4881	0.4458	0.3847	0.0775	0.1277	0.1942	0.3902	0.3787	0.3783	0.3441	0.5774
	DFR BM25	0.4780	0.4407	0.3839	0.0764	0.1256	0.1934	0.3843	0.3749	0.3765	0.3428	0.5763
	TF-IDF	0.5288	0.4695	0.4110	0.0860	0.1388	0.2079	0.4278	0.4088	0.4061	0.3578	0.5938
	Dirichlet LM	0.4576	0.4220	0.3661	0.0691	0.1121	0.1707	0.3585	0.3571	0.3549	0.3120	0.5567
	Hiemstra LM	0.5424	0.4915	0.4297	0.0862	0.1408	0.2149	0.4328	0.4183	0.4200	0.3671	0.6001
Remove hyphens	BM25	0.5322	0.4966	0.4407	0.0869	0.1437	0.2265	0.4315	0.4270	0.4324	0.4042	0.6570
	DFR BM25	0.5288	0.4949	0.4407	0.0860	0.1425	0.2250	0.4259	0.4247	0.4302	0.4026	0.6547
	TF-IDF	0.5966	0.5390	0.4686	0.0942	0.1580	0.2425	0.4812	0.4677	0.4649	0.4224	0.6783
	Dirichlet LM	0.5051	0.4729	0.4153	0.0813	0.1338	0.2034	0.3961	0.3967	0.4014	0.3729	0.6364
	Hiemstra LM	0.5797	0.5542	0.4907	0.0946	0.1620	0.2503	0.4688	0.4726	0.4771	0.4338	0.6827
Remove accents	BM25	0.4915	0.4441	0.3754	0.0798	0.1283	0.1900	0.3977	0.3815	0.3773	0.3413	0.5798
	DFR BM25	0.4780	0.4373	0.3746	0.0784	0.1261	0.1889	0.3899	0.3778	0.3754	0.3396	0.5785
	TF-IDF	0.5153	0.4627	0.3966	0.0870	0.1369	0.2013	0.4280	0.4074	0.4011	0.3540	0.5948
	Dirichlet LM	0.4542	0.4102	0.3534	0.0657	0.1055	0.1592	0.3517	0.3457	0.3421	0.3037	0.5497
	Hiemstra LM	0.5424	0.4881	0.4186	0.0861	0.1405	0.2108	0.4382	0.4198	0.4153	0.3634	0.6001
Remove stopwords	BM25	0.5390	0.4847	0.4212	0.0891	0.1465	0.2182	0.4286	0.4157	0.4156	0.3716	0.6098
	DFR BM25	0.5356	0.4864	0.4195	0.0886	0.1463	0.2169	0.4267	0.4155	0.4138	0.3706	0.6088
	TF-IDF	0.5559	0.4831	0.4203	0.0924	0.1470	0.2193	0.4423	0.4185	0.4174	0.3716	0.6123
	Dirichlet LM	0.4712	0.4271	0.3703	0.0723	0.1176	0.1788	0.3640	0.3553	0.3585	0.3184	0.5656
	Hiemstra LM	0.5458	0.5102	0.4347	0.0895	0.1505	0.2243	0.4341	0.4283	0.4263	0.3796	0.6164
Heavy stemming	BM25	0.4881	0.4695	0.3992	0.0819	0.1409	0.2096	0.3944	0.3970	0.3949	0.3650	0.6060
	DFR BM25	0.4881	0.4712	0.3983	0.0818	0.1410	0.2088	0.3941	0.3979	0.3939	0.3643	0.6056
	TF-IDF	0.5525	0.4949	0.4339	0.0925	0.1500	0.2252	0.4449	0.4270	0.4298	0.3810	0.6244
	Dirichlet LM	0.4712	0.4441	0.3805	0.0701	0.1197	0.1810	0.3563	0.3616	0.3631	0.3290	0.5762
	Hiemstra LM	0.5458	0.5153	0.4492	0.0868	0.1479	0.2275	0.4312	0.4321	0.4366	0.3871	0.6261
Moderate stemming	BM25	0.4915	0.4712	0.3992	0.0822	0.1413	0.2099	0.3970	0.3981	0.3953	0.3656	0.6065
	DFR BM25	0.4915	0.4712	0.3983	0.0821	0.1411	0.2090	0.3967	0.3982	0.3942	0.3648	0.6060
	TF-IDF	0.5525	0.4949	0.4339	0.0925	0.1500	0.2252	0.4449	0.4270	0.4298	0.3814	0.6246
	Dirichlet LM	0.4712	0.4441	0.3805	0.0701	0.1197	0.1810	0.3563	0.3616	0.3631	0.3293	0.5764
	Hiemstra LM	0.5458	0.5153	0.4492	0.0868	0.1479	0.2276	0.4312	0.4321	0.4366	0.3874	0.6263
Light stemming	BM25	0.4949	0.4763	0.4008	0.0826	0.1421	0.2106	0.3996	0.4015	0.3963	0.3675	0.6079
	DFR BM25	0.4949	0.4763	0.4008	0.0825	0.1420	0.2102	0.3993	0.4016	0.3958	0.3668	0.6075
	TF-IDF	0.5559	0.5000	0.4381	0.0929	0.1508	0.2266	0.4475	0.4305	0.4323	0.3837	0.6261
	Dirichlet LM	0.4814	0.4508	0.3831	0.0716	0.1216	0.1829	0.3655	0.3701	0.3691	0.3317	0.5817
	Hiemstra LM	0.5492	0.5203	0.4542	0.0880	0.1496	0.2305	0.4360	0.4369	0.4421	0.3906	0.6301

The results presented in Table 7.5 were further analyzed by combining text preprocessing techniques to better understand their collective impact on long-text retrieval effectiveness. These combined preprocessing techniques yielded varied impacts across evaluation metrics and models. Hiemstra LM consistently outperformed other models in most combinations and cutoffs, except in P@5 and NDCG@5, where TF-IDF exhibited a slight performance advantage. Notable results for Hiemstra LM include the combination of apostrophe and hyphen removal, which achieved the

highest scores at P@10 (0.5576), P@20 (0.4907), MAP@20 (0.2523), and NDCG@20 (0.4790).

Additional configurations, such as incorporating *moderate* stemming, yielded the best results at MAP@5 (0.1645), while including stopwords removal improved NDCG@10 (0.4760) and overall MAP (0.4358). Furthermore, Hiemstra LM showed notable gains in effectiveness when combining hyphen and accent removal, particularly at MAP@10 (0.1645) and NDCG (0.6855). Hyphen removal alone continued to demonstrate its effectiveness in long-text retrieval, achieving the same highest score as its combination with apostrophe removal at P@20 (0.4907).

Table 7.6: Effectiveness of combined text preprocessing techniques in long-text retrieval. Values highlighted with a *green* background indicate the best score at the respective cutoff.

Retrieval Strategies	Model	Precision at Cutoff			MAP at Cutoff			NDCG at Cutoff			MAP	NDCG
		@5	@10	@20	@5	@10	@20	@5	@10	@20		
Baseline	BM25	0.4847	0.4525	0.3839	0.0769	0.1281	0.1931	0.3883	0.3800	0.3765	0.3429	0.5764
	DFR BM25	0.4746	0.4475	0.3839	0.0758	0.1259	0.1925	0.3826	0.3763	0.3753	0.3416	0.5754
	TF-IDF	0.5288	0.4746	0.4110	0.0855	0.1382	0.2068	0.4275	0.4086	0.4045	0.3564	0.5927
	Dirichlet LM	0.4576	0.4186	0.3669	0.0696	0.1111	0.1701	0.3577	0.3544	0.3545	0.3110	0.5558
	Hiemstra LM	0.5390	0.4915	0.4314	0.0856	0.1402	0.2138	0.4289	0.4158	0.4189	0.3655	0.5990
Remove hyphens	BM25	0.5322	0.4966	0.4407	0.0869	0.1437	0.2265	0.4315	0.4270	0.4324	0.4042	0.6570
	DFR BM25	0.5288	0.4949	0.4407	0.0860	0.1425	0.2250	0.4259	0.4247	0.4302	0.4026	0.6547
	TF-IDF	0.5966	0.5390	0.4686	0.0942	0.1580	0.2425	0.4812	0.4677	0.4649	0.4224	0.6783
	Dirichlet LM	0.5051	0.4729	0.4153	0.0813	0.1338	0.2034	0.3961	0.3967	0.4014	0.3729	0.6364
	Hiemstra LM	0.5797	0.5542	0.4907	0.0946	0.1620	0.2503	0.4688	0.4726	0.4771	0.4338	0.6827
Remove apostrophes and hyphens	BM25	0.5322	0.4949	0.4407	0.0869	0.1434	0.2271	0.4300	0.4254	0.4328	0.4047	0.6567
	DFR BM25	0.5254	0.4932	0.4407	0.0857	0.1423	0.2256	0.4230	0.4232	0.4307	0.4031	0.6543
	TF-IDF	0.5966	0.5390	0.4678	0.0944	0.1581	0.2432	0.4788	0.4664	0.4650	0.4232	0.6790
	Dirichlet LM	0.5017	0.4729	0.4136	0.0809	0.1348	0.2043	0.3953	0.3983	0.4011	0.3737	0.6369
	Hiemstra LM	0.5831	0.5576	0.4907	0.0955	0.1636	0.2523	0.4721	0.4759	0.4790	0.4355	0.6840
Remove hyphens and accents	BM25	0.5322	0.4932	0.4280	0.0875	0.1442	0.2222	0.4349	0.4251	0.4273	0.4020	0.6585
	DFR BM25	0.5254	0.4932	0.4280	0.0861	0.1439	0.2211	0.4316	0.4272	0.4278	0.4007	0.6578
	TF-IDF	0.5831	0.5288	0.4525	0.0948	0.1541	0.2339	0.4744	0.4588	0.4555	0.4175	0.6769
	Dirichlet LM	0.5017	0.4576	0.4025	0.0779	0.1261	0.1923	0.3882	0.3847	0.3875	0.3645	0.6288
	Hiemstra LM	0.5831	0.5508	0.4763	0.0955	0.1645	0.2474	0.4812	0.4748	0.4737	0.4330	0.6855
Remove hyphens and stopwords	BM25	0.5797	0.5220	0.4619	0.0926	0.1547	0.2361	0.4614	0.4492	0.4521	0.4195	0.6722
	DFR BM25	0.5763	0.5203	0.4568	0.0918	0.1549	0.2344	0.4585	0.4483	0.4491	0.4189	0.6718
	TF-IDF	0.6102	0.5339	0.4661	0.0961	0.1574	0.2403	0.4849	0.4622	0.4612	0.4232	0.6777
	Dirichlet LM	0.4983	0.4695	0.4068	0.0742	0.1276	0.1946	0.3841	0.3860	0.3886	0.3604	0.6235
	Hiemstra LM	0.5932	0.5458	0.4797	0.0947	0.1601	0.2451	0.4780	0.4722	0.4714	0.4339	0.6840
Remove hyphens, apostrophes, and stopwords	BM25	0.5729	0.5153	0.4610	0.0922	0.1540	0.2366	0.4575	0.4457	0.4514	0.4199	0.6726
	DFR BM25	0.5695	0.5136	0.4576	0.0913	0.1542	0.2352	0.4545	0.4447	0.4492	0.4193	0.6720
	TF-IDF	0.6102	0.5322	0.4661	0.0960	0.1570	0.2410	0.4838	0.4618	0.4608	0.4236	0.6779
	Dirichlet LM	0.4881	0.4678	0.4042	0.0731	0.1270	0.1940	0.3786	0.3848	0.3858	0.3611	0.6232
	Hiemstra LM	0.6000	0.5492	0.4797	0.0959	0.1628	0.2471	0.4826	0.4760	0.4727	0.4358	0.6853
Remove hyphens, apostrophes, and apply moderate stemmer	BM25	0.5322	0.5017	0.4322	0.0849	0.1459	0.2236	0.4156	0.4176	0.4202	0.3981	0.6450
	DFR BM25	0.5254	0.4966	0.4297	0.0844	0.1447	0.2214	0.4110	0.4136	0.4173	0.3968	0.6435
	TF-IDF	0.6000	0.5441	0.4636	0.0942	0.1583	0.2364	0.4681	0.4589	0.4561	0.4160	0.6700
	Dirichlet LM	0.4915	0.4712	0.4076	0.0788	0.1330	0.1989	0.3795	0.3872	0.3911	0.3663	0.6274
	Hiemstra LM	0.5864	0.5559	0.4805	0.0965	0.1631	0.2475	0.4695	0.4734	0.4702	0.4301	0.6780
Average relative performance gains of the best model compared to the baseline		15.39%	13.45%	13.75%	12.73%	17.33%	18.01%	13.43%	14.48%	14.35%	19.23%	14.44%
Average relative performance gains of the best model compared to others within the same retrieval strategy		9.61%	11.77%	11.47%	13.23%	16.40%	12.54%	9.64%	10.15%	11.08%	7.82%	4.65%

In contrast, individual preprocessing techniques, such as accent and apostrophe removal, generally underperformed when applied in isolation, contributing minimally to retrieval effectiveness across different metrics (see Table 7.5). However, combining these techniques with hyphen and stopwords removal, along with *moderate* stemming, resulted in a strong performance, particularly with Hiemstra LM across most metrics and cutoffs. While the *heavy* stemming variant slightly outperformed the *moderate* variant when applied individually, the *moderate* variant showed a slight advantage when combined with other preprocessing techniques.

Hiemstra LM consistently delivered the highest scores with the combined preprocessing approach across all metrics, except at P@5, where TF-IDF marginally outperformed it. Hiemstra LM demonstrated notable performance improvements over the baseline, achieving gains of up to 13.75% for Precision cutoffs, 18.01% for MAP cutoffs, 14.48% for NDCG cutoffs, 19.23% for overall MAP, and 14.44% for overall NDCG. Additionally, Hiemstra LM achieved an average relative improvement of up to 16.40% over the other retrieval models. Conversely, BM25 and DFR BM25 did not produce the best results with any preprocessing strategies for long-text retrieval, suggesting that BM25-based models may be less effective in handling long-document contexts. These findings underscore the different impacts of text preprocessing techniques and retrieval models on short- and long-text retrieval tasks in Tetun.

7.2.7 Discussion

In Tetun ad-hoc text retrieval, text preprocessing techniques such as the removal of apostrophes and hyphens generally result in performance improvements, with relative gains of up to 8.19% observed using DFR BM25 in short-text retrieval up to the top-10 cutoffs. When stopword removal is combined with apostrophe and hyphen removal, further gains are observed, particularly in MAP@5 with both DFR BM25 and TF-IDF models and in NDCG with Hiemstra LM. Notably, when apostrophes are retained and only hyphens and stopwords are removed, Hiemstra LM demonstrates relative improvements of up to 8.33% over other models at P@20, MAP@20, NDCG@20, and overall MAP. These findings suggest that targeting hyphen removal significantly enhances retrieval effectiveness for short-text retrieval.

Different strategies for handling stopwords and apostrophes yield varied benefits across retrieval models. Retaining stopwords while removing apostrophes proves more effective with DFR BM25, whereas retaining apostrophes while removing stopwords shows better results with Hiemstra LM at higher cutoffs and overall. Accent removal, however, does not enhance retrieval effectiveness in short-text retrieval, whether applied individually or in combination with other preprocessing techniques.

Removing stopwords does not lead to noticeable improvements in retrieval effectiveness. For instance, at P@20, the relative gain is limited to 0.58% compared to Hiemstra LM with apostrophe and hyphen removal. Similarly, MAP@20 improves by 0.76%, NDCG@20 by 0.07%, and overall MAP by 1.22% relative to the baselines. These findings align with Savoy's [199] experiments with French, which showed that retaining stopwords yielded better results with the BM25 model. Similar conclusions were reported by Ghosh and Bhattacharya [100], who found variability in retrieval effectiveness within the same language and no significant improvement from stopword removal. Additionally, Dolamic and Savoy [81] observed that retaining stopwords improved retrieval effectiveness in Hindi, with minimal differences for Marathi and Bengali.

Regarding stemming, it does not improve retrieval effectiveness in Tetun, whether applied independently or in combination with other preprocessing techniques. This result is consistent with findings from preliminary experiments on Tetun text retrieval [61] and Flores and Moreira's [93] experiments with Dutch, Italian, Spanish, and English. However, studies on other LRLs, including

those by Sahu and Pal [191] (Sanskrit), Sahu et al. [192] (Urdu), and Adriani et al. [2] (Indonesian), reported positive impacts of stemming on retrieval effectiveness. This discrepancy likely arises from Tetun’s language-specific characteristics, which may respond differently to stemming than other languages.

While intrinsic evaluations suggested that the *moderate* and *heavy* stemmers were slightly more accurate than the *light* stemmer, these differences did not consistently translate into improved retrieval performance. These findings align with Flores and Moreira’s [93] observation that higher stemmer accuracy does not always enhance retrieval effectiveness.

For long-text retrieval, Hiemstra LM consistently outperformed other models across most metrics and cutoffs. Hyphen removal delivered performance gains across various metrics and models, particularly when combined with apostrophe removal, achieving high scores at P@10, P@20, MAP@20, and NDCG@20. Adding moderate stemming to this combination yielded the highest MAP@5 score, though the improvement was minimal compared to strategies without stemming.

Apostrophe removal showed inconsistent results as an individual preprocessing technique but became more effective when combined with hyphen removal. Similarly, accent removal had varying impacts across models and metrics when applied alone, but combined with other techniques, it achieved a relative improvement of 16.40% in MAP@10 over other models. This outcome aligns with Savoy’s [199] findings on French, where accent removal slightly improved retrieval effectiveness in long-text retrieval but negatively affected short-text retrieval for several strategies.

7.2.8 Conclusions

The most effective retrieval strategy for Tetun ad-hoc text retrieval involves short-text retrieval (i.e., using document titles) combined with targeted preprocessing techniques, specifically splitting compound words by removing hyphens and eliminating apostrophes from queries and documents. Together, these techniques enhance both retrieval efficiency and effectiveness. The removal of stopwords shows minimal impact, while preserving accents proves crucial for effective text retrieval in Tetun, likely due to the language’s reliance on diacritics to distinguish word meanings.

Stemming does not improve retrieval effectiveness in short-text retrieval. In long-text retrieval, the *moderate* stemming variant, when combined with other preprocessing techniques, shows slight improvements at MAP@5; however, the gains remain marginal and are still lower than the best-performing score achieved in short-text retrieval, representing -39.25% relative performance drop. This minimal impact suggests that morphological normalization through stemming is not essential for effective text retrieval in Tetun, likely due to the language’s relatively simple morphology and minimal use of inflectional affixes.

The experimental results show that text preprocessing techniques, particularly the removal of hyphens and apostrophes, are crucial in enhancing the effectiveness of Tetun ad-hoc text retrieval. Among the retrieval models, DFR BM25 performs best for cutoffs up to the top-10 but exhibits greater sensitivity to stopword removal. In contrast, Hiemstra LM provides the most substantial improvements at top-20 cutoffs and beyond, particularly in overall MAP and NDCG.

Overall, this study demonstrates the importance of thoroughly investigating language-specific preprocessing strategies by segmenting them into distinct stages and systematically integrating them to uncover the linguistic components that impact retrieval effectiveness (see [Figure 7.1](#)). It emphasizes the benefit of developing approaches tailored to the unique characteristics of each language, rather than relying solely on established techniques in the literature. The study introduces a detailed and adaptable methodology for establishing ad-hoc text retrieval baselines, particularly for LRLs where such benchmarks have not yet been established. By accounting for the unique morphological and syntactic features of each language, researchers can design more effective and linguistically appropriate retrieval strategies.

7.3 Search Prototype

To showcase the most effective retrieval strategies from the baseline for Tetun ad-hoc text retrieval, we developed a search prototype using the Django framework,² which handles backend operations, including application logic, request processing, and interface rendering. The frontend was designed using Django’s built-in Jinja2 templating engine and styled with Bootstrap 3,³ ensuring a responsive and user-friendly interface. Documents from the Labadain-30k+ dataset were stored in a PostgreSQL database to support efficient data storage and management.

For document retrieval, the documents on the PostgreSQL database were systematically indexed in Solr using the Django Haystack API,⁴ which serves as a connecting layer between Django and Solr, streamlining the integration of indexing and search functionalities. The prototype interface is illustrated in [Figure 7.2](#).

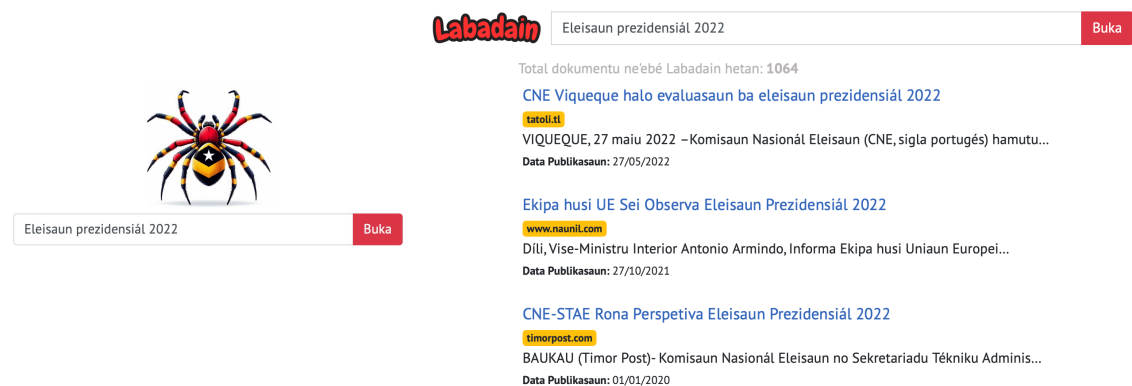


Figure 7.2: Tetun ad-hoc text retrieval prototype.

Prior to indexing, preprocessing steps, including lowercasing and apostrophe removal, were applied using Solr’s built-in functions. Tokenization was performed using Solr’s standard tokenizer, which also splits hyphenated words. Document titles were indexed, and BM25 was employed as the retrieval and ranking model with default settings. While DFR BM25 was not directly

²<https://www.djangoproject.com>

³<https://getbootstrap.com>

⁴<https://django-haystack.readthedocs.io/en/master/>

supported in this setup, experimental results showed only marginal performance differences between BM25 and DFR BM25, making BM25 a practical and effective choice for this retrieval system prototype.

The search prototype provides a user-friendly platform for querying and retrieving documents in Tetun. It features a search box for users to input their queries and a results page displaying documents ranked by relevance. Each page displays up to 20 results, with each result including the document title, a clickable URL linking to the source, a snippet of the content, and the publication date. The system also supports functionality for navigating across multiple result pages.

The prototype has been hosted online and accessible at <https://www.labadain.tl> since 26 January 2024, allowing users to explore and interact with the system effectively. This implementation demonstrates the practical application of retrieval models and preprocessing strategies optimized for Tetun text, providing a foundation for future advancements in Tetun text retrieval systems. In 2024, the system recorded an average of 27 unique visitors per month, reflecting its limited public visibility and the infrequent updates of the indexed documents. The total number of unique visitors per month is presented in Figure 7.3.

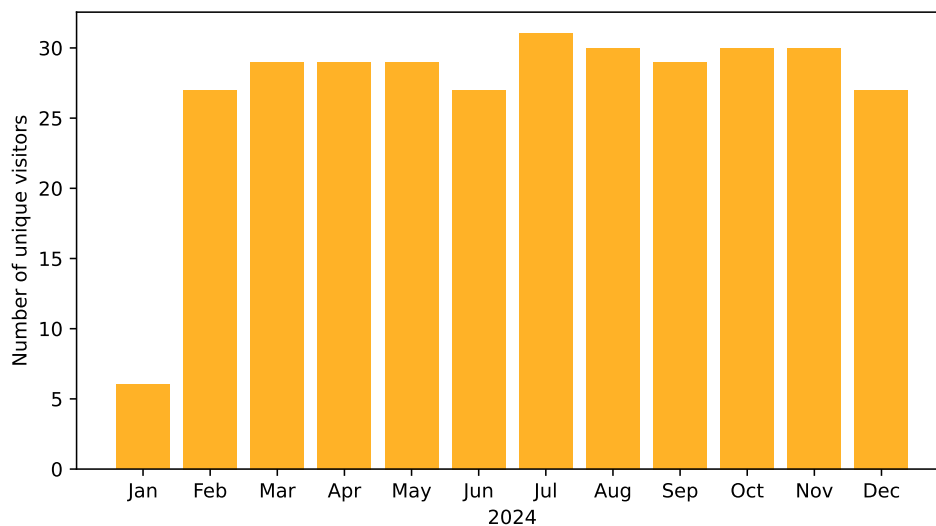


Figure 7.3: Total unique visitors per month in 2024.

7.4 Zero-Shot and Hybrid Retrieval

Traditional sparse retrieval models, such as BM25, which rely on lexical matching, have demonstrated strong performance in many IR systems [54, 138, 188]. However, because these models retrieve and rank documents based on exact term matching, they are prone to vocabulary mismatch and fail to capture semantic relationships between queries and documents [24, 55, 105, 126, 161].

To bridge this gap, transformer-based dense retrieval models, such as BERT [79] and its variants, have been widely used in text retrieval tasks [90, 126, 141, 145, 167, 234, 243]. These models encode queries and documents independently into low-dimensional vector representations, where

semantic relevance is estimated based on vector similarity, typically computed using cosine similarity or dot product functions [90, 105, 145].

Dense retrieval models are generally trained using supervised learning approaches for representation learning, which require a labeled dataset (i.e., query-document pairs). However, training a retrieval model from scratch is not feasible for most languages, particularly those that are under-resourced, due to data scarcity and computational constraints. As an alternative, pretrained dense retrieval models can be fine-tuned for specific tasks or applied directly in zero-shot settings.

Given the limited availability of labeled data for Tetun and the fact that existing pretrained dense retrieval models do not currently support the language, this section presents results from an investigation into the use of dense retrievers in zero-shot, out-of-distribution scenarios. The models were adapted to Tetun documents to produce zero-shot embeddings, and their performance was evaluated across different document representations and retrieval strategies.

7.4.1 Overview

Training and fine-tuning dense retrieval models for text retrieval remain computationally expensive and typically require large-scale labeled datasets, which are often difficult to obtain for IR tasks [129]. As a promising alternative, zero-shot text retrieval has gained increasing attention in recent years [39, 185, 129, 216, 249]. More recently, researchers have begun exploring the potential of large language models (LLMs) to support retrieval tasks [123, 255].

MS MARCO [18, 165] is a widely used dataset for training and fine-tuning dense retrieval models and generally uses BERT or mBERT as the backbone models for training. Several dense retrieval models have been trained or fine-tuned on this dataset, including mDPR [249], Contriever and mContriever [119], ColBERT [128, 196], mColBERT [27], ColBERT-X [140], ColBERT-XM [149], ANCE [241], and coCondenser [97, 141].

Studies demonstrate that hybrid retrieval—which combines a traditional sparse retriever with a dense retriever—consistently outperforms a sparse retriever alone, regardless of whether dense retrieval models are fine-tuned [98, 126, 141, 153] or used in zero-shot scenarios, both in- and out-of-distributions [39, 249, 250] and in high- and low-resource languages (LRLs) [249, 250, 251]. Two hybrid retrieval approaches have been proposed: (1) score fusion through a linear combination of lexical and semantic matching [98, 126, 141, 153, 249, 251] and (2) reciprocal rank fusion (RRF) [39, 48].

Given the promising results of zero-shot out-of-distribution retrieval in both high- and low-resource languages [39, 249, 250], this work investigates its feasibility for Tetun, a language not currently supported by existing pretrained dense retrieval models, which makes the setting especially challenging. We explore the application of a zero-shot out-of-distribution approach to Tetun, investigating the following research questions (RQs):

RQ1: How effective are zero-shot monolingual and multilingual dense retrievers when applied to Tetun in out-of-distribution scenarios compared to traditional sparse retrievers?

RQ2: What is the impact of combining sparse and dense retrievers on the zero-shot retrieval effectiveness for Tetun?

RQ3: How does augmenting document titles with LLM-generated summaries impact zero-shot retrieval effectiveness in Tetun?

To address these RQs, we investigate the performance of pretrained dense retrieval models under zero-shot and hybrid retrieval strategies in out-of-distribution scenarios. Given prior findings that classical sparse retrievers in Tetun perform more effectively when using document titles (see Section 7.2), and that dense retrievers generally benefit from richer contextual signals, we hypothesize that augmenting document titles with content summaries and encoding these enriched representations with a pretrained dense retrieval model may lead to improved retrieval effectiveness for Tetun. We use summaries instead of full documents due to the input length limitations of dense retrievers, which make processing long texts impractical. This approach also allows us to explore how well an LLM can generate content summaries in Tetun.

To test this hypothesis, we begin by selecting pretrained dense retrieval models that have demonstrated strong performance in text retrieval tasks. Subsequently, these models are applied to Tetun queries and document titles (step 1 in Figure 7.4), and to contextual documents constructed by concatenating document titles with summaries generated by an LLM (step 2 in Figure 7.4), to create zero-shot embeddings. Next, query and document vectors are computed using cosine similarity to determine their relevance. Finally, the results are linearly combined to produce the final retrieval output.

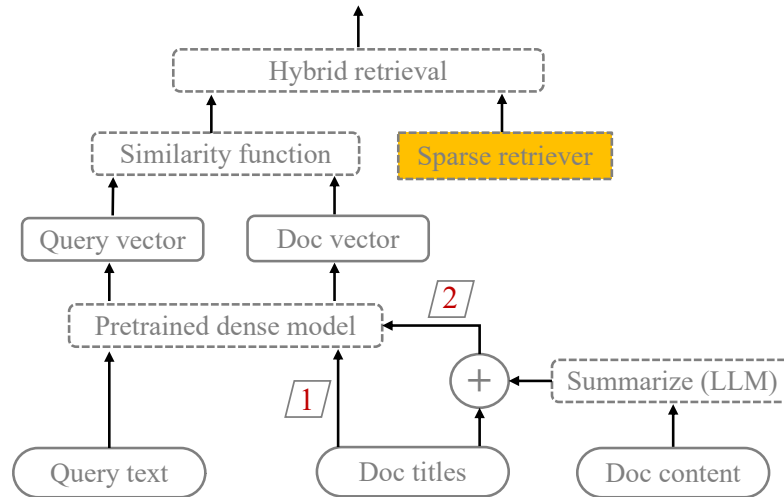


Figure 7.4: Overview of the hybrid retrieval strategy.

To evaluate the effectiveness of dense retrieval models for Tetun, we compare zero-shot retrieval performance against the established baseline. Specifically, we assess the performance of

dense retrievers, hybrid retrieval using document titles, and hybrid retrieval enhanced with contextual information. The results show that hybrid retrieval performs best when document titles are enriched with summaries generated by an LLM from the documents' content, based on evaluations of up to 2,000 retrieved documents per query.

7.4.2 Related Work

Pretrained dense retrieval models have shown growing potential for improving retrieval effectiveness across languages and retrieval strategies. Early work, such as dense passage retrieval (DPR) [126], demonstrated the effectiveness of dual-encoder architectures in open-domain question answering. This approach was later extended to multilingual settings through mDPR [249], which introduced Mr. TyDI, a multilingual benchmark for monolingual text retrieval with a particular emphasis on low-resource languages.

Khattab and Zaharia [128] introduced ColBERT, a dense retrieval model that enables fine-grained interaction between queries and documents by comparing token-level representations, allowing for efficient and effective relevance estimation. Its successor, ColBERTv2 [196], improved both accuracy and scalability through re-ranking and compression mechanisms. The multilingual extensions mColBERT [27], ColBERT-XM [149], and ColBERT-X [140] further adapted the architecture to support retrieval across languages, including low-resource ones. Additionally, models such as ANCE [241], which uses iterative negative sampling, and Contriever and mContriever [119], and coCondenser [97, 141], which are trained without supervision, further advance dense models for text retrieval.

In zero-shot retrieval scenarios, Thakur et al. [216] reported that BM25 maintained strong performance across various corpora, while dense retrievers generally underperformed relative to BM25. Only ColBERT achieved superior results on the BEIR benchmark, outperforming BM25 by an average of +2.5% in NDCG@10 during first-stage retrieval. Similarly, Zhang et al. [249] investigated monolingual ad-hoc text retrieval in non-English languages, focusing on out-of-domain scenarios by training mDPR with mBERT [239] as a replacement for DPR's original encoder. Their findings showed that mDPR did not outperform BM25 when evaluated on the Mr. TyDi benchmark. Building on Mr. TyDi, Zhang et al. [250] introduced MIRACL and reported that dense retrievers, such as mDPR and mContriever, continued to underperform BM25 in zero-shot, out-of-domain settings when evaluated using NDCG@10.

For hybrid retrieval, Karpukhin et al. [126] proposed a score fusion method in which the top 2,000 passages retrieved independently by sparse (BM25) and dense (DPR) retrievers are merged, and their relevance scores are combined using a linear weighted sum. A tuning parameter (λ) was applied to the dense retriever to control its contribution in the fusion process. Experiments on several question-answering (QA) datasets demonstrated that setting $\lambda = 1.1$ on the development set improved end-to-end QA match accuracy. This approach was later adapted by Ma et al. [153] using the same datasets, but the λ tuning parameter was applied to the sparse retriever instead. Its value was optimized via grid search over the range [0, 2] with a step size of 0.05, yielding comparable results.

Gao et al. [98] also experimented with a hybrid retrieval approach for first-stage ad-hoc text retrieval, in which the dense retriever was fine-tuned on the MS MARCO passage dataset [18]. In this setup, the λ parameter was applied to the sparse retriever during testing, yielding significant improvements over sparse retrieval alone when evaluated on both the MS MARCO and TREC 2019 Deep Learning [52] evaluation query sets. Likewise, Lee et al. [141] evaluated score fusion on MS MARCO passages and TREC Robust04 [230], placing the λ parameter on the dense retriever. The optimal value was selected based on MAP performance on a development set, using a grid search over the range $[0, 2]$ with a step size of 0.1. They found optimal parameter values $\lambda = 1.5$ for MS MARCO and $\lambda = 2.0$ for Robust04. With this configuration, their hybrid approach outperformed existing baselines by combining BM25 with their trained coCondenser model. In a related study, Zhang et al. [250] evaluated hybrid retrieval on the MIRACL datasets, applying α to the sparse retriever and $(1 - \alpha)$ to the dense retriever, with α fixed at 0.5 without tuning. Their results revealed that combining mDPR with BM25 outperformed individual retrievers in most languages.

In zero-shot, out-of-domain settings, Zhang et al. [249] evaluated the generalizability and robustness of dense retrievers using a hybrid strategy, where a tuning parameter α was applied to the dense retriever. The value of α was optimized over the range $[0, 1]$ via simple line search on the development set, using MRR@100 as the objective and a step size of 0.01. Combining the top 1,000 results from mDPR and BM25, the hybrid approach significantly outperformed BM25 alone in nine out of eleven languages when evaluated using MRR@100 and Recall@100.

Motivated by these findings, we explore two key directions for extension in Tetun. First, while prior work typically applies a single tuning parameter to the sparse or dense retriever, we assume that assigning independent parameters to each may more effectively balance their contributions, particularly since sparse and dense models often assign different relevance scores to the same document for a given query. Second, whereas previous studies have primarily focused on passage retrieval and have not incorporated contextual information, we explore the use of enriched document representations. Specifically, we augment document titles with summaries generated by an LLM from the documents' content, and adapt pretrained dense retrievers to these contextualized inputs to generate zero-shot embeddings.

7.4.3 Methodology

To assess the performance of pretrained dense retrieval models in Tetun, we define four key retrieval settings by leveraging existing pretrained models for text retrieval and adapting them to Tetun documents to produce zero-shot embeddings. We then evaluate how these models perform under different retrieval configurations and examine the impact of contextual augmentation and hybrid retrieval approaches.

The study investigates two key factors: *document representation* and *retrieval strategy*. The former determines whether documents are represented using only titles or are augmented with summaries generated by an LLM to provide additional contextual information. The latter examines

whether retrieval relies solely on dense models or integrates dense and sparse models within a hybrid framework. Based on these factors, four retrieval settings are defined as follows:

1. **Basic Zero-Shot Dense Retrieval:** Pretrained dense retrieval models are applied independently to Tetun queries and document titles to produce zero-shot embeddings, and their retrieval performance is evaluated.
2. **Contextualized Zero-Shot Dense Retrieval:** Building on (1), document titles are augmented with text summaries generated by LLMs to create embeddings. This setting assesses whether enriched contextual information improves zero-shot dense retrieval performance.
3. **Basic Hybrid Retrieval:** The results from the zero-shot dense retrieval models in (1) are combined with those from sparse retrieval models that use only document titles. This setting examines the effectiveness of hybrid retrieval without additional contextual information.
4. **Contextualized Hybrid Retrieval:** Building on (3), this strategy combines results from the contextual dense retrieval models in (2) with sparse retrieval outputs to assess the impact of enriched contextual information in a hybrid setup.

We first examine the performance of various pretrained dense retrieval models when adapted to Tetun text documents to produce zero-shot embeddings. We then investigate the impact of incorporating contextual information to enhance these embeddings. Next, we evaluate hybrid retrieval techniques by combining sparse and dense retrievers (see Figure 7.4). Additionally, we compare the performance of monolingual and multilingual variants of the same model.

As shown in Table 7.4 in Section 7.2, the baseline results indicate that BM25, DFR BM25, and Hiemstra LM demonstrated comparable performance across various cutoff points. Therefore, all three classical models were included in the hybrid retrieval experiments, and their performance was evaluated using P@10, MAP@10, and NDCG@10.

We focus on pretrained dense retrieval models that have demonstrated strong performance in text retrieval tasks across diverse domains. Selection criteria include the availability of both monolingual and multilingual variants (particularly with support for LRLs), public accessibility, and prior application in hybrid retrieval scenarios. The models selected are DPR [126], mDPR [249], Contriever and mContriever [119], ColBERT [128, 196], and ColBERT-X [140].

7.4.4 Dual Parameters for Hybrid Retrieval

In hybrid retrieval, score fusion typically involves combining the scores from both the sparse and dense retrievers through a linear weighted sum, while assigning a tuning parameter to adjust the contribution of either the sparse or dense retriever. Prior studies have explored various single-parameter tuning strategies, focusing on adjusting either the sparse retriever [98, 153], the dense retriever [126, 141, 249, 251], or both using a single shared parameter [250].

Given that sparse and dense retrievers often assign different relevance scores to the same document for a given query, we assign independent parameters to each retriever to balance their contributions more effectively. Following established conventions for linear combination [98, 126, 141, 249, 153, 250], we define the dual-parameter tuning approach as follows:

$$\text{Score}(q, d) = \alpha \cdot \text{Score}_{\text{lex}}(q, d) + \beta \cdot \text{Score}_{\text{sem}}(q, d) \quad (7.1)$$

To identify the optimal combination of α and β , we adopt a strategy inspired by Khramtsova et al. [129] for selecting dense retrievers in zero-shot search scenarios. We perform a grid search over predefined ranges of α and β , evaluating each search step across multiple retrieval strategies and metrics using *qrels* to identify optimal parameter values. At each step of the search, only the parameter values that yield the highest score for each retrieval strategy and metric are recorded for comparison. Given the strong performance of sparse baselines for Tetun, we set α to start from 1.

7.4.5 Experimental Setup

The experimental configurations comprise document summarization and retrieval settings, each of which is described in detail below.

7.4.5.1 Document Summarization and Zero-Shot Embedding Generation

For document summarization, the document content was fed into the LLM to generate a summary. The Haiku variant of the Claude 3 model from Anthropic⁵ was used for this task, selected for its cost-effectiveness and its potential to support text summarization in Tetun [65]. The user and system prompts used to instruct the LLM are presented in Prompt 7.4.1.

These prompts were developed based on preliminary pilot experiments, during which we qualitatively assessed several prompt variants by observing the outputs to identify those that produced coherent and representative summaries of the Tetun documents' content. The LLM was instructed to generate exactly one paragraph of summary for each input document. To guide the model, an example input document in Tetun and its corresponding expected summary were provided in each LLM call. To produce zero-shot embeddings for Tetun, we adapted pretrained dense retrieval models to encode queries and documents. For the documents, we explored two types of representations: titles alone (step 1 in Figure 7.4) and contextual documents that augment titles with summaries generated from the document content using an LLM (step 2 in Figure 7.4).

7.4.5.2 Zero-Shot and Hybrid Retrieval Strategies

Two strategies were tested and evaluated for zero-shot retrieval. The first strategy involves calculating query-document relevance by computing the cosine similarity between query embeddings and document title embeddings. The second uses contextual embeddings, titles concatenated with content summaries, comparing them to the query embeddings using the same similarity function.

⁵<https://www.anthropic.com>

The performance of each selected pretrained dense retrieval model, as outlined in Subsection 7.4.3, was evaluated for both strategies.

For zero-shot hybrid retrieval, the document scores produced by each of the zero-shot dense retrievers were linearly combined with the scores from one of the three sparse retrievers (BM25, DFR BM25, and Hiemstra LM), using the approach proposed in Subsection 7.4.4. To determine the optimal values of α and β , we performed a grid search over $\alpha \in [1, 2]$ and $\beta \in [0, 2]$ with step sizes of 0.01, 0.05, and 0.1, and evaluated performance using P@10, MAP@10, and NDCG@10.

Prompt 7.4.1: Details of the LLM Prompt.

User prompt:

Provide a concise contextual description that summarizes the following Tetun document. Do not include any additional text or generate hallucinated content. Do not respond in bullet points, and write exactly one paragraph.

{document}

Follow this example:

Input document in Tetun:

{document_example}

Expected summary:

{summary_example}

System prompt:

You are an expert in Tetun and text understanding. Assume the document is entirely in Tetun and respond using accurate Tetun grammar. Keep your response in Tetun, including correct spelling, punctuation, and other linguistic aspects. Ignore incomplete sentences.

Additionally, the models that outperformed the baseline were further evaluated using the single-parameter linear combination approach and the reciprocal rank fusion (RRF) method to compare their performance. Finally, the performance of monolingual and multilingual dense retrieval models adapted to Tetun was assessed.

7.4.6 Results and Analysis

This section presents the experimental results and analysis for six dense retrieval models—DPR, mDPR, Contriever, mContriever, ColBERTv2, and ColBERT-X—evaluated under two configurations: *basic* (using document titles only) and *contextualized* (using document titles enriched with LLM-generated summaries). Performance is evaluated across two retrieval strategies: zero-shot retrieval, in which Tetun documents are adapted to generate zero-shot embeddings, and hybrid

retrieval, which combines sparse and dense signals through linear score fusion and reciprocal rank fusion, using three evaluation metrics: P@10, MAP@10, and NDCG@10.

7.4.6.1 Zero-Shot Retrieval

Figure 7.5 shows the relative performance gains of zero-shot dense retrieval models compared with baseline sparse retrievers in both *basic* and *contextualized* settings. Overall, all dense retrieval models underperform the sparse baseline in zero-shot Tetun text retrieval. Moreover, their performance generally declines further when contextual information is introduced, as illustrated by the orange bars in Figure 7.5.

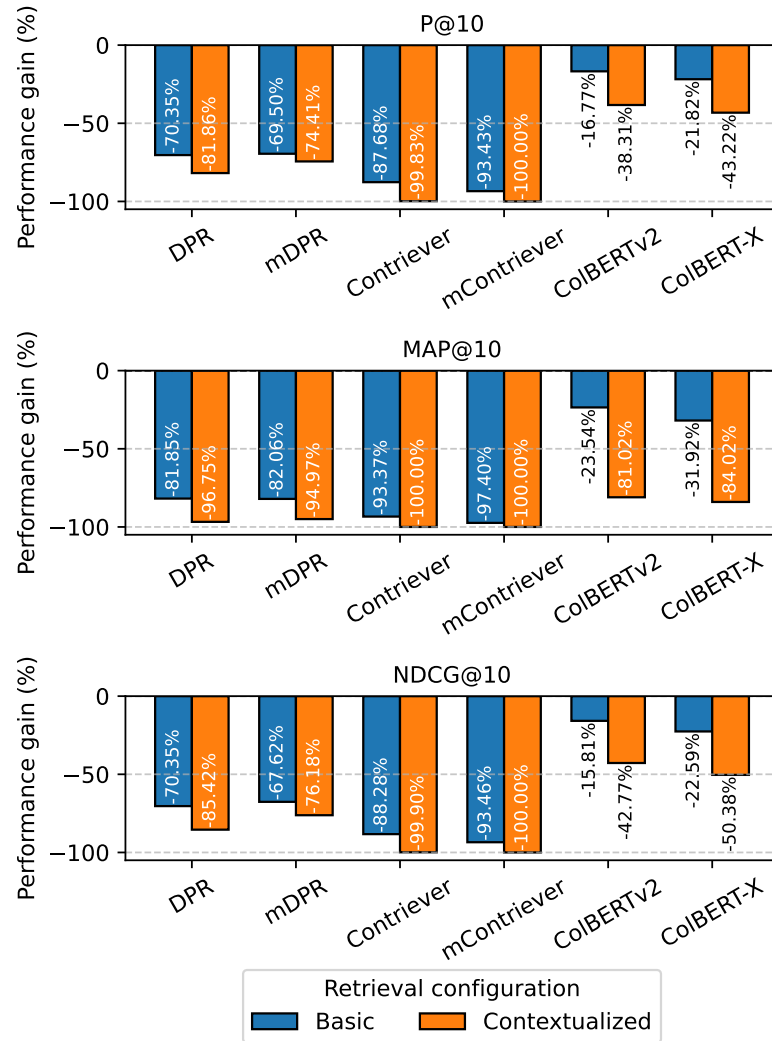


Figure 7.5: Relative performance gains of zero-shot dense retrieval over the baseline.

Among the evaluated models, Contriever and mContriever perform the weakest across all metrics, with most of the configurations showing performance losses reaching -100% when contextual information is introduced. DPR and mDPR also underperform consistently relative to the sparse baseline, with further degradation of up to -15% when contextualized inputs are used. While

ColBERTv2 and ColBERT-X do not surpass the sparse baseline, they achieve the strongest performance among the dense retrievers under the basic configuration. However, their effectiveness declines considerably with the introduction of enriched document representations, with performance drops of up to -50%.

When comparing each monolingual dense retrieval model to its multilingual counterpart, mixed trends are observed. None of the multilingual models was trained on Tetun. For the DPR family, the multilingual variant exhibits a smaller performance drop than its monolingual counterpart. In contrast, for both Contriever and ColBERT variants, the monolingual models demonstrate lower performance loss compared to their multilingual counterparts.

7.4.6.2 Basic Hybrid Retrieval

In hybrid retrieval with *basic* configuration, we conducted a grid search over $\alpha \in [1, 2]$ and $\beta \in [0, 2]$, and the optimal parameters identified were $\alpha = 1.0$ and $\beta = 0.3$, with a step size of 0.1. The relative performance gains over the DFR BM25 baseline for this strategy are shown in [Figure 7.6](#).

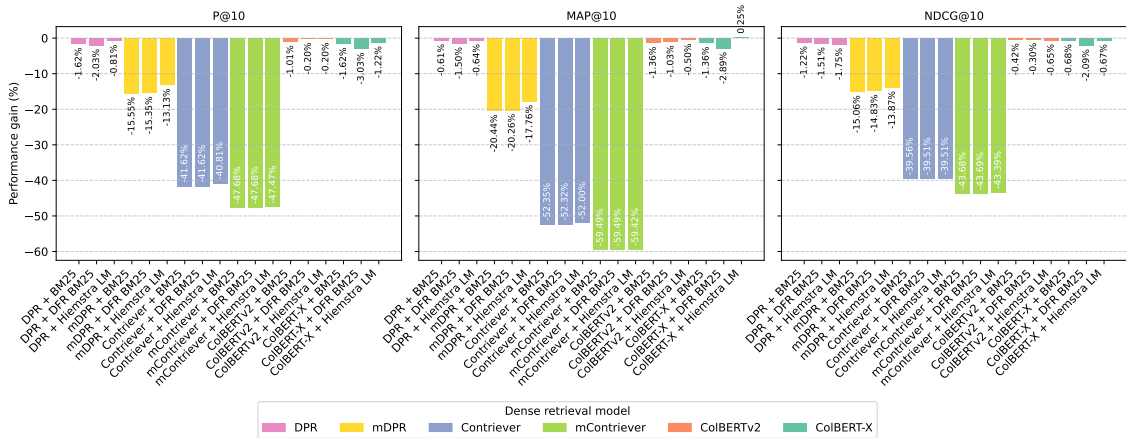


Figure 7.6: Relative performance gains of hybrid retrieval over the baseline under the *basic* configuration ($\alpha = 1.0$, $\beta = 0.3$).

Overall, most combinations of dense retrievers with Hiemstra LM yield the smallest performance losses relative to the baseline. Notably, the ColBERT-X + Hiemstra LM combination slightly outperforms the baseline, achieving a +0.25% relative performance gain in MAP@10 (middle bar graph in [Figure 7.6](#)). In contrast, the Contriever family continues to exhibit the largest performance degradation, with losses reaching up to -60%. Among the evaluated models, DPR, ColBERTv2, and ColBERT-X demonstrate minimal performance loss when combined with sparse retrievers, each under 5% compared to the baseline. In contrast, mDPR shows moderate degradation, with losses ranging from -13% to -20%. Contriever continues to perform significantly worse, with drops between -40% and -52%, while mContriever records the highest performance degradation, ranging from -47% to -60%.

When comparing *basic* hybrid configuration results to their zero-shot retrieval counterparts shown in [Figure 7.5](#), significant performance improvements are observed. For instance, at MAP@10,

and when combined with Hiemstra LM, DPR improves from a -81.85% loss in the zero-shot setting to just -0.64% ; Contriever improves from -93.37% to -52.00% ; and most notably, ColBERT-X shifts from a -84.02% performance drop in zero-shot to a $+0.25\%$ gain under this hybrid setup.

Regarding the comparison between monolingual and multilingual dense retrieval models, nearly all monolingual models outperform their multilingual counterparts when combined with sparse retrievers and are generally close to the baseline performance. An exception is ColBERT-X combined with Hiemstra LM at MAP@10, where it slightly exceeds the baseline with a relative performance gain of $+0.25\%$. ColBERTv2, meanwhile, exhibits consistent performance across all combinations and evaluation metrics.

7.4.6.3 Contextualized Hybrid Retrieval

Based on the optimal parameter values identified through grid search in the *basic* hybrid retrieval setup, where $\alpha = 1.0$ indicated that no adjustment to the sparse retriever and optimal performance was achieved using a single tuning parameter $\beta = 0.3$ for the dense retriever, we extended the evaluation of single-parameter tuning to the *contextualized* configuration.

In this regard, we assess both single-parameter and dual-parameter fusion strategies: single-parameter tuning applies either α to the sparse retriever or β to the dense retriever, while the dual-parameter setup assigns independent weights to both. Moreover, we also evaluate reciprocal rank fusion (RRF) as an additional fusion method. The results for contextualized hybrid retrieval are presented in Table 7.7. Given our objective to demonstrate that balancing signal fusion from both dense and sparse retrievers, particularly when incorporating contextual information, yields better performance than one-sided tuning, we begin by presenting the results from the dual-parameter configuration, followed by the single-parameter settings.

Dual Parameters The results of dual-parameter tuning, with optimal parameter values of $\alpha = 1.1$ and $\beta = 0.8$, are presented in Table 7.7. An example illustrating the optimal α and β values obtained from the grid search, which yielded the highest MAP@10 score, is shown in Figure 7.7. The results show that combining ColBERTv2 with each of the sparse retrievers consistently outperforms the baseline across all evaluation metrics. The best performance is achieved by the ColBERTv2 + Hiemstra LM combination, yielding a relative improvement of $+2.01\%$ in P@10, $+4.24\%$ in MAP@10, and $+2.45\%$ in NDCG@10 over the baseline, based on evaluations of up to 2,000 retrieved documents per query. These improvements are statistically significant, as confirmed by a paired t -test ($p < 0.05$). Scores exceeding the baseline are shown in bold, with the best results highlighted in green. Conversely, other dense retrievers exhibit less consistent behavior when introducing contextual information. For instance, when combined with Hiemstra LM in MAP@10, ColBERT-X incurs only a marginal additional loss of -0.14 percentage points relative to its *basic* hybrid configuration. In contrast, DPR, mDPR, Contriever, and mContriever experience substantially greater degradations, with additional losses ranging from -4.67 to -25.57 percentage points.

Table 7.7: Performance of contextualized hybrid retrieval in Tetun. Scores in **bold** indicate improvement over the baseline. Values with a *green* background represent the best results ($p < 0.05$). The symbol † indicates that the model outperformed its counterpart (monolingual vs. multilingual).

	P@10	MAP@10	NDCG@10
Baseline			
DFR BM25	0.8390	0.2804	0.7356
Contextualized hybrid retrieval ($\alpha = 1.1$ and $\beta = 0.8$, top 2k results)			
DPR + BM25	0.6864 (-18.19%)	0.2079 (-25.86%)	0.6161 (-16.25%)
DPR + DFR BM25	0.6864 (-18.19%)	0.2083 (-25.71%)	0.6169 (-16.14%)
DPR + Hiemstra LM	0.6847 (-18.39%)	0.2061 (-26.50%)	0.6102 (-17.05%)
mDPR + BM25	0.7068 (-15.76%) †	0.2226 (-20.61%) †	0.6236 (-15.23%) †
mDPR + DFR BM25	0.7102 (-15.35%) †	0.2234 (-20.33%) †	0.6262 (-14.87%) †
mDPR + Hiemstra LM	0.7288 (-13.13%) †	0.2305 (-17.80%) †	0.6334 (-13.89%) †
Contriever + BM25	0.3169 (-62.23%) †	0.0791 (-71.79%) †	0.2938 (-60.06%) †
Contriever + DFR BM25	0.3169 (-62.23%) †	0.0791 (-71.79%) †	0.2947 (-59.94%) †
Contriever + Hiemstra LM	0.3220 (-61.62%) †	0.0796 (-71.61%) †	0.2985 (-59.42%) †
mContriever + BM25	0.1814 (-78.38%)	0.0408 (-85.45%)	0.1936 (-73.68%)
mContriever + DFR BM25	0.1814 (-78.38%)	0.0409 (-85.41%)	0.1938 (-73.65%)
mContriever + Hiemstra LM	0.1847 (-77.99%)	0.0421 (-84.99%)	0.1967 (-73.26%)
ColBERTv2 + BM25	0.8441 (+0.61%)†	0.2862 (+2.07%)†	0.7453 (+1.32%)†
ColBERTv2 + DFR BM25	0.8458 (+0.81%)†	0.2858 (+1.93%)†	0.7480 (+1.69%)†
ColBERTv2 + Hiemstra LM	0.8559 (+2.01%)†	0.2923 (+4.24%)†	0.7536 (+2.45%)†
ColBERT-X + BM25	0.8305 (-1.01%)	0.2765 (-1.39%)	0.7335 (-0.29%)
ColBERT-X + DFR BM25	0.8322 (-0.81%)	0.2779 (-0.89%)	0.7308 (-0.65%)
ColBERT-X + Hiemstra LM	0.8356 (-0.41%)	0.2807 (+0.11%)	0.7303 (-0.72%)
Contextualized hybrid retrieval (single parameter, $\alpha = 1.1$, top 2k results)			
ColBERTv2 + BM25	0.8508 (+1.41%)†	0.2888 (+3.00%)†	0.7493 (+1.86%)†
ColBERTv2 + DFR BM25	0.8373 (-0.20%) †	0.2831 (+0.96%)†	0.7438 (+1.11%)†
ColBERTv2 + Hiemstra LM	0.8475 (+1.01%)†	0.2891 (+3.10%)†	0.7504 (+2.01%)†
ColBERT-X + BM25	0.8322 (-0.81%)	0.2771 (-1.18%)	0.7334 (-0.30%)
ColBERT-X + DFR BM25	0.8339 (-0.61%)	0.2771 (-1.18%)	0.7325 (-0.42%)
ColBERT-X + Hiemstra LM	0.8356 (-0.41%)	0.2790 (-0.50%)	0.7304 (-0.71%)
Contextualized hybrid retrieval (single parameter, $\beta = 0.8$, top 2k results)			
ColBERTv2 + BM25	0.8458 (+0.81%)†	0.2866 (+2.21%)†	0.7467 (+1.51%)†
ColBERTv2 + DFR BM25	0.8441 (+0.61%)†	0.2856 (+1.85%)†	0.7476 (+1.63%)†
ColBERTv2 + Hiemstra LM	0.8542 (+1.81%)†	0.2917 (+4.03%)†	0.7530 (+2.37%)†
ColBERT-X + BM25	0.8322 (-0.81%)	0.2775 (-1.03%)	0.7344 (-0.16%)
ColBERT-X + DFR BM25	0.8339 (-0.61%)	0.2774 (-1.07%)	0.7327 (-0.39%)
ColBERT-X + Hiemstra LM	0.8373 (-0.20%)	0.2804 (+0.00%)	0.7305 (-0.69%)
Reciprocal Rank Fusion			
ColBERTv2 + BM25	0.7949 (-5.26%) †	0.2641 (-5.81%) †	0.7164 (-2.61%) †
ColBERTv2 + DFR BM25	0.7966 (-5.05%) †	0.2642 (-5.78%) †	0.7166 (-2.58%) †
ColBERTv2 + Hiemstra LM	0.8220 (-2.03%) †	0.2738 (-2.35%) †	0.7316 (-0.54%) †
ColBERT-X + BM25	0.7898 (-5.86%)	0.2575 (-8.17%)	0.6965 (-5.32%)
ColBERT-X + DFR BM25	0.7898 (-5.86%)	0.2577 (-8.10%)	0.6969 (-5.26%)
ColBERT-X + Hiemstra LM	0.8034 (-4.24%)	0.2604 (-7.13%)	0.6985 (-5.04%)

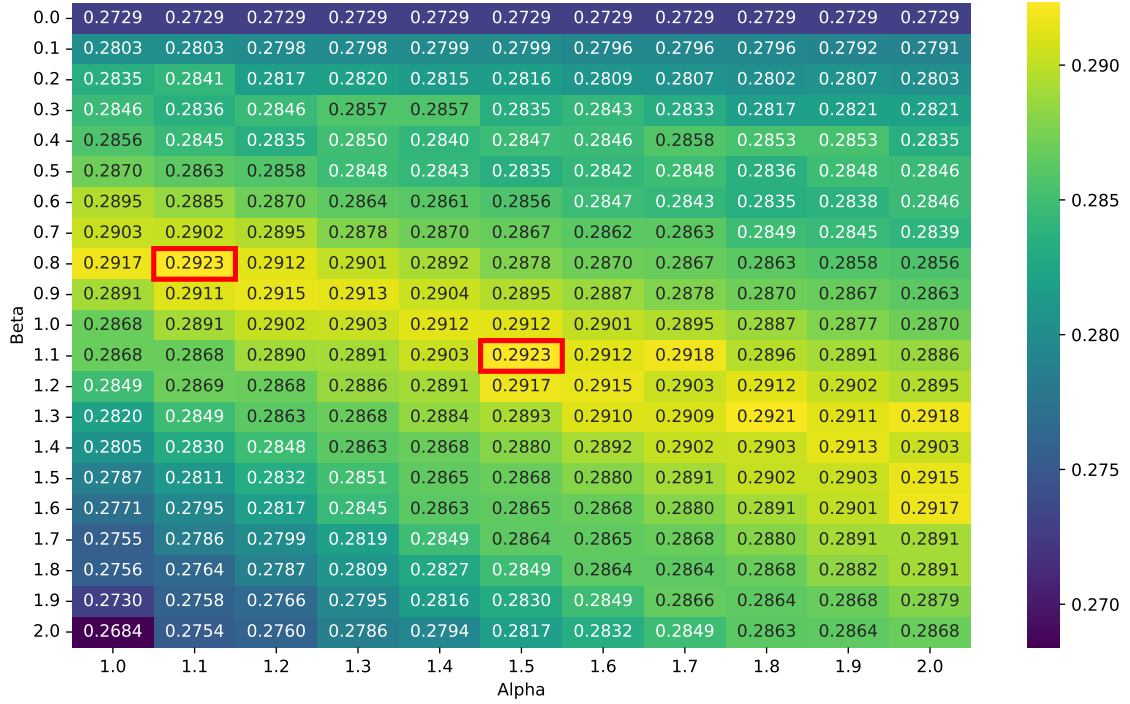


Figure 7.7: Grid search results for the *contextualized* configuration combining ColBERTv2 with the Hiemstra LM, evaluated using MAP@10. The combinations $\alpha = 1.1$, $\beta = 0.8$ and $\alpha = 1.5$, $\beta = 1.1$ yield the highest MAP@10 score of 0.2923 (highlighted with red boxes). Since only the top-scoring combination is recorded, the first one encountered during the grid search is selected.

Single Parameter To evaluate the effectiveness of single-parameter tuning, we selected the models that outperformed the baseline in the dual-parameter setting while preserving the optimal parameter values identified for the sparse and dense retrievers. Thus, ColBERTv2 and ColBERT-X were chosen for this analysis, with the corresponding results reported in Table 7.7. Overall, tuning a single parameter for the ColBERTv2-based model, either in the sparse or dense retriever, continues to yield strong performance relative to the baseline, except for the ColBERTv2 + DFR BM25 combination when the parameter is applied to the sparse retriever. In contrast, ColBERT-X-based combinations show less consistent behavior. For example, the previously observed MAP@10 improvement when combined with Hiemstra LM drops to -0.50% in the single-parameter setting. In general, applying the tuning parameter to the dense retriever proves more effective than tuning the sparse retriever, except in the case of ColBERTv2 combined with BM25. Notably, tuning the sparse retriever in the ColBERTv2 + BM25 combination outperforms its dual-parameter counterpart across all metrics, achieving relative gains of up to +0.93 percentage points.

Reciprocal Rank Fusion To further assess the effectiveness of the hybrid retrieval strategy for Tetun ad-hoc text retrieval, we evaluated the performance of ColBERTv2 and ColBERT-X using the reciprocal rank fusion (RRF) method, as shown in Table 7.7. While the combination of ColBERTv2 + Hiemstra LM shows a slight advantage, the overall results suggest that RRF does not

outperform the baseline and offers limited benefit for its application to Tetun. In the *contextualized* hybrid configuration, nearly all monolingual models consistently outperform their multilingual counterparts. The only exception is the mDPR model, which achieves better results across all evaluation metrics in the dual-parameter setting.

7.4.7 Discussion

The results of this study are consistent with prior studies [216, 249, 250], showing that all evaluated dense retrieval models underperform relative to the sparse baseline (DFR BM25) in zero-shot scenarios, regardless of whether the models are monolingual or multilingual, with monolingual models exhibiting a slight advantage over their multilingual counterparts. These outcomes reaffirm the limitations of dense retrieval models in generalizing across domains and languages outside their training distribution, particularly in LRLs, and highlight the robustness of sparse models.

A contextualization strategy introduced in this study—enriching document representations by concatenating document titles with LLM-generated content summaries—does not improve performance in zero-shot, out-of-distribution scenarios. On the contrary, it leads to further performance degradation, as illustrated in Figure 7.5 (orange bars). These findings suggest that current dense retrieval models cannot leverage additional contextual signals in unfamiliar domains and may be particularly sensitive to noisy, redundant, or domain-mismatched inputs.

Hybrid retrieval, in contrast, when dense retrievers were combined with sparse models under the *basic* hybrid configuration (see Figure 7.6), all evaluated combinations showed notable gains relative to their zero-shot counterparts in Figure 7.5. The hybrid configuration combining ColBERT-X with Hiemstra LM was particularly promising, achieving near-baseline performance and, in one case, slightly surpassing it by +0.25% in MAP@10. While these findings are broadly consistent with prior work, they differ in input data: previous studies predominantly employ passages for training or fine-tuning, whereas this study uses document titles, which are significantly shorter and often less informative. This distinction likely contributes to the lower standalone effectiveness of dense retrievers in our setting. Nevertheless, the results highlight the efficacy of combining lexical and semantic signals, even when the dense model underperforms in isolation.

Further improvements were observed in the *contextualized* hybrid setting using dual-parameter tuning, as presented in Table 7.7. ColBERTv2, when combined with Hiemstra LM and tuned with $\alpha = 1.1$ and $\beta = 0.8$, yielded statistically significant gains across all evaluation metrics (P@10, MAP@10, and NDCG@10), with relative improvements of up to +4.24%. This suggests that properly calibrating the balance between sparse and dense signals is essential to maximizing hybrid retrieval effectiveness. Interestingly, these improvements were not uniformly observed across all models, reinforcing that architectural differences among dense retrievers impact their ability to leverage contextual and fused signals.

Single-parameter tuning experiments further reveal that strong performance can still be achieved when tuning only one component. For ColBERTv2, sparse-side tuning in combination with BM25 not only performed competitively but also outperformed the model’s dual-parameter configuration by up to +0.93 percentage points. This finding suggests that in practical scenarios, particularly

those that rely on BM25 as the sparse retriever, a simpler fusion strategy focused on the sparse side may offer an effective and efficient alternative to more complex dual-parameter approaches.

The performance of reciprocal rank fusion (RRF), a rank-based hybridization technique, was less consistent. Although some benefit was observed for ColBERTv2 combined with Hiemstra LM, RRF generally did not outperform score fusion methods and often underperformed relative to the baseline. This outcome indicates that, in the context of Tetun ad-hoc text retrieval, score-based hybridization provides finer control over the influence of each retriever and may be more suitable than rank-only fusion strategies.

The comparison between monolingual and multilingual dense retrievers revealed a notable performance gap favoring monolingual models, particularly in hybrid configurations. This suggests that multilingual models may not generalize Tetun effectively without further adaptation. An exception was ColBERT-X, which, despite its multilingual nature, demonstrated strong stability and nearly matched or exceeded the performance of monolingual ColBERTv2 in certain configurations. This indicates that multilingual dense retrievers can be competitive when paired with late interaction mechanisms and carefully tuned hybrid strategies.

The relationship between Tetun’s linguistic characteristics, the model’s pretraining, and the retrieval architecture is a key factor influencing retrieval performance. Tetun’s typological status, as a language with extensive lexical borrowing from Portuguese [62, 103, 106, 223, 224], is reflected in the query and document content, with an average of approximately 47% of tokens in queries and 30% in contextual documents being Portuguese loanwords, as shown in Table 7.8. Examples of such loanwords in Tetun include “*ekonomia*” (from “*economia*”, meaning “economy”) and “*progresu*” (from “*progresso*”, meaning “progress”). This linguistic overlap may help explain the robust performance of ColBERT-X, although it does not fully account for the consistently higher performance of monolingual ColBERTv2. One plausible explanation, aligned with the broader discussion on the balance between sparse and dense signals, is that both the queries and the documents favor models that excel at lexical matching.

Table 7.8: Statistics for queries and documents containing Portuguese loanwords. Loanwords were identified by matching each preprocessed word against the Portuguese loanword dictionary published by Greksáková [103]. *Tokens include words and numbers.

Description	Min	Max	Avg
Total tokens (queries)*	3	5	3.46
Total loanwords (queries)	0	3	1.61
Total tokens (contextual documents)	4	335	122.87
Total loanwords (contextual documents)	0	99	36.57

From this perspective, the competitive performance of the ColBERT architecture may stem from its late-interaction mechanism, which preserves token-level signals. Likewise, the superior performance of the monolingual ColBERT variant might be attributed to the high proportion of Portuguese (Latinate) loanwords in Tetun, which aligns well with ColBERTv2’s pretraining on English, a language that also contains substantial Latinate influence in its lexicon and morphology.

Overall, these findings emphasize the nuanced interplay between retrieval architecture, fusion strategy, and contextual input. While dense retrievers face substantial challenges in zero-shot Tetun ad-hoc text retrieval, their performance can be substantially enhanced through hybrid approaches, architectural choices, and tuning parameter balancing. The results reinforce attention to hybrid retrieval design and suggest contextual adaptation in LRL scenarios.

7.4.8 Conclusions

This study investigates the retrieval effectiveness of adapting pretrained dense retrievers to Tetun, a low-resource language (LRL), under zero-shot and hybrid retrieval scenarios in out-of-distribution settings. In the zero-shot setting, in response to RQ1, our results confirm that dense retrievers, both monolingual and multilingual, consistently underperform relative to the sparse retriever baseline. This performance gap highlights the limitations of current dense retrieval models in generalizing to LRLs not seen during training. While monolingual dense models slightly outperform their multilingual counterparts, none surpass the baseline, reaffirming the robustness of traditional sparse retrieval models in low-resource, out-of-domain contexts and underscoring the challenge of adapting dense models for such scenarios.

In hybrid retrieval configurations, addressing RQ2 and RQ3, our findings show that hybrid approaches are generally less effective when using short text inputs, such as document titles. However, performance improves substantially when these titles are enriched with contextual information. Specifically, augmenting titles with summaries generated by a large language model (LLM) before encoding leads to notable gains. The best performance is achieved with the ColBERTv2 + Hiemstra LM combination. A relative improvement in MAP@10 of +3.10% is observed when assigning $\alpha = 1.1$ to the sparse retriever, and +4.03% when assigning $\beta = 0.8$ to the dense retriever. The highest improvement of +4.24% in MAP@10 is achieved when applying both parameters ($\alpha = 1.1, \beta = 0.8$), indicating the effectiveness of the dual-tuning approach in hybrid retrieval.

These improvements highlight the effectiveness of both single and dual-parameter tuning strategies when contextual information is incorporated. This supports a contextualization approach that is particularly beneficial in scenarios where passage-level datasets are unavailable, a common challenge in under-resourced language settings. The enhanced retrieval effectiveness observed through contextual enrichment confirms our hypothesis that augmenting short-text representations with LLM-generated summaries can significantly improve zero-shot retrieval performance.

This work introduces two methodological contributions. First, it applies a dual-parameter tuning strategy, motivated by the observation that sparse and dense retrievers often assign different relevance scores to the same document. By weighting each retriever independently, this method balances their complementary signals more effectively, leading to improved retrieval performance. Second, it proposes a contextualized retrieval strategy for zero-shot scenarios, in which LLM-generated summaries are concatenated with document titles to enrich representations before generating dense embeddings. To support reproducibility and foster future research, we publicly release two resources [76] under the Creative Commons Attribution Share-Alike (CC BY-SA) license: contextual document summaries generated by an LLM and retrieval run files.

Chapter 8

Tetun Users’ Search Behavior Analysis

Understanding the search behavior of Tetun-speaking users is essential for identifying their information needs and designing effective, user-centered search systems. Although query log analysis has been widely studied, no prior work has specifically examined the behavior of Tetun-speaking users in the context of either conventional search engines or large language model (LLM)-based conversational assistants.

This chapter presents an analysis of Tetun-speaking users’ search behavior, based on research conducted as part of this thesis [125]. The analysis focuses on search prompts submitted to an LLM-based conversational assistant and includes a comparative perspective with studies involving English-speaking users in similar contexts. To support future research, we also release LabadainLog-17k+ [71], a dataset of more than 17,000 real-world Tetun user search logs. This resource provides a unique opportunity to investigate conversational search behavior in Tetun.

8.1 Overview

Conversational search refers to user-friendly dialogues between humans and machines, whether spoken or written, aimed at satisfying information needs [12, 14, 182, 245]. With the advent of LLM-based conversational assistants, such as OpenAI’s ChatGPT,¹ Google’s Gemini,² Anthropic’s Claude,³ and similar AI tools, such dialogues have become more natural and context-aware, closely resembling human conversation. These systems’ ability to process unstructured language, handle complex queries, and support multi-turn interactions has made search more seamless and intuitive.

Furthermore, these developments demand a deeper understanding of user behavior, providing deeper insights into search behavior, engagement, and intent. These insights are crucial for refining LLM-based conversational systems to provide more personalized and effective interactions, ultimately enhancing the search experience [253, 143]. Despite these advancements, limited access to interaction log data hinders understanding of how these systems are used [218]. This challenge is

¹<https://chatgpt.com>

²<https://gemini.google.com/app>

³<https://claude.ai>

especially pronounced in underrepresented language communities, such as Tetun speakers, where such data remain unavailable.

To address these challenges, we launched Labadain Chat in early 2024,⁴ an LLM-based conversational assistant for Tetun speakers, aimed at studying search behavior, engagement, and intent in the Tetun-speaking community. The platform's home page is illustrated in Figure 8.1. The system was built using a combination of LangChain,⁵ Streamlit,⁶ and the Haiku variant of Claude 3 from Anthropic, with Haiku selected for its low cost and strong performance with Tetun.



Figure 8.1: Labadain Chat home page.

For analysis, we collect user prompt logs, IP addresses, and user agents under established terms and conditions, with users acknowledging that their data are collected for research purposes. Both the user prompts and responses are recorded in a text file automatically generated for each day. As of July 2024, registration became mandatory for platform access, enabling the collection of more data for user interaction analysis.

We analyzed 16,952 interactions recorded between July 22 and October 3, 2024, from 904 unique users. The analysis addresses the following research questions (RQs):

RQ1. *What similarities and differences can be observed when comparing interaction logs of Tetun- and English-speaking users of LLM-based conversational assistants?*

RQ2. *How have LLM-based conversational assistants influenced user search trends on traditional search platforms?*

To investigate RQ1, we analyze Labadain Chat prompt logs to examine search behaviors, engagement, and intent among Tetun-speaking users, focusing on user growth, usage frequency,

⁴<https://www.labadain.com>

⁵<https://www.langchain.com>

⁶<https://streamlit.io>

session duration, and topics of interest. We compare these findings with Google Bard logs for English-speaking users, as reported by Trippas et al. [218], to identify similarities and differences in interaction logs and search behaviors. Additionally, we analyze follow-up search sequences based on themes from Kim et al. [130] to gain further insights.

To address RQ2, we compare search behaviors, intent, and prompt quality across three datasets: Labadain Chat [68] interaction logs, Google Search Console query logs (capturing searches leading to clicks on Timor News [59] links),⁷ and Labadain Search [69] logs, derived from a search engine prototype developed for Tetun speakers.

The results reveal that Tetun-speaking users exhibit search characteristics similar to those of English-speaking users. Additionally, LLM-based conversational systems influence user search behavior in traditional search engines, reflecting a trend toward the use of command-based instructions in queries.

8.2 Related Work

Conversational search aims to satisfy information needs through user-friendly dialogues, either in spoken or written form [12, 14, 182]. The Dagstuhl Seminar 19461 report [12] underscores the need to understand user search behaviors to enhance human-machine dialogues for information retrieval. Zamani et al. [245] recently emphasized that conversational systems should assist users in completing tasks by retrieving relevant information.

The analysis of user query logs is crucial for understanding user behavior, intent, and interaction patterns in both traditional search engines and LLM-based conversational assistants. Query log analysis has been extensively studied for conventional search engines, covering areas such as query intent distribution [190], the evolution of search behaviors on specific topics [84], re-finding previously accessed search results [219], identifying interaction patterns [202], and analyzing traffic from specific topical categories [23]. In contrast, research on LLM-based conversational assistant logs remains limited due to restricted access to interaction log data [218], and no studies have analyzed interaction log data in Tetun.

In a recent study, Trippas et al. [218] analyzed Google Bard’s user interaction logs for English, comprising 15,023 interactions (14,595 prompts) from 95 unique crowd worker respondents. To the best of our knowledge, no other study of this nature has been conducted to date. The findings revealed that users primarily engaged in short sessions, with nearly half consisting of only one interaction. Function words dominated unigram and bigram frequencies, and user intents were predominantly command-driven. These prompts frequently included requests for tasks such as text extraction, summarization, text formatting, and information retrieval. Additionally, three main user intent categories were identified: *imperatives* (commands or requests), *modal or auxiliary verbs* (questions or expressions of possibility), and *pronouns* (referencing subjects in questions).

Few public datasets exist for LLM-based conversational search, limiting our understanding of their usage. To our knowledge, only two publicly available datasets exist for English: *BIDD-1k*,

⁷<https://search.google.com/search-console/about>

which contains 1,000 interaction prompts collected from crowd workers [218], and *LMSYS-Chat-1M*, comprising one million Chatbot Arena query logs [253]. However, no studies have analyzed Chatbot Arena data in the context of user search behaviors. Given the lack of such data in Tetun, this study addresses the gap by releasing LabadainLog-17k+ [71], the first publicly available LLM-based real-world user search log dataset in the language.

8.3 Platforms and Datasets

The following subsections provide an overview of the platforms from which the data were collected and detail the data collection processes for each.

8.3.1 Platforms Overview

The log data used in this study were derived from three different platforms: Labadain Chat, Labadain Search, and Google Search Console for Timor News. Labadain Chat [68] is an LLM-based conversational system developed specifically for Tetun speakers and launched in January 2024. In July 2024, registration became mandatory to access the platform, with an average of 43 unique active users per day recorded through early October 2024.

Labadain Search [69], a monolingual search engine tailored for Tetun speakers, was launched in the same month as Labadain Chat and averaged 55 unique visitors per day. Timor News [59] is an online news portal exclusively publishing content in Tetun. Operational since May 2019, it recorded an average of 873 unique visitors per day. The portal was integrated with Google Search Console to log queries from Google searches when users clicked links directing them to the site.

8.3.2 Data Collection and Datasets

Labadain Chat records user prompts and their corresponding responses, along with metadata such as IP addresses and user agents. We analyzed 16,952 user prompts recorded between July 22, 2024, and October 3, 2024, from 904 unique users, representing 76.48% of registered users during that period. A sample of the interaction log, with English translations, is shown in Table 8.1.

Table 8.1: Sample of interaction logs from Labadain Chat.

2024-07-22 09:38:50 103.94.XX.XXX Mozilla/5.0... username	=> Dadeer di’ak Labadain!
Translation:	Good morning Labadain!
2024-08-22 08:25:25 Labadain	=> Dadeer di’ak! Ha’u kontente tebes atu simu ita-boot. Ha’u-nia naran mak Labadain, asistente AI ida hosi Timor News. Ha’u sei responde ita-nia pergunta sira hotu uza Tetun ofisiál. Oinsá ha’u bele tulun ita-boot?
Translation:	Good morning! I’m very happy to welcome you. My name is Labadain, an AI assistant from Timor News. I will answer all your questions using the official Tetun. How can I assist you?

For Labadain Search, the system captures anonymous user session IDs along with the request timestamp and query text. During the same timeframe as the Labadain Chat data collection, we gathered 605 query logs. Additionally, we collected the top 1,000 query logs from Google Search Console during this period, selecting the highest-ranked queries based on click count.

8.4 Methodology

An overview of the methodology is shown in Figure 8.2. The analysis began with Labadain Chat prompt logs, which were examined to identify user search behavior and interaction patterns, including temporal usage trends. These logs were subsequently analyzed to extract engagement metrics, providing insights into the platform’s user base. In the final stage, prompt logs from Labadain Chat were compared with search query logs from both Labadain Search and Google Search Console for Timor News. This cross-platform analysis aimed to investigate user search intent, topical distributions, query reformulation sequences, and behavioral evolution across conversational and traditional search systems.

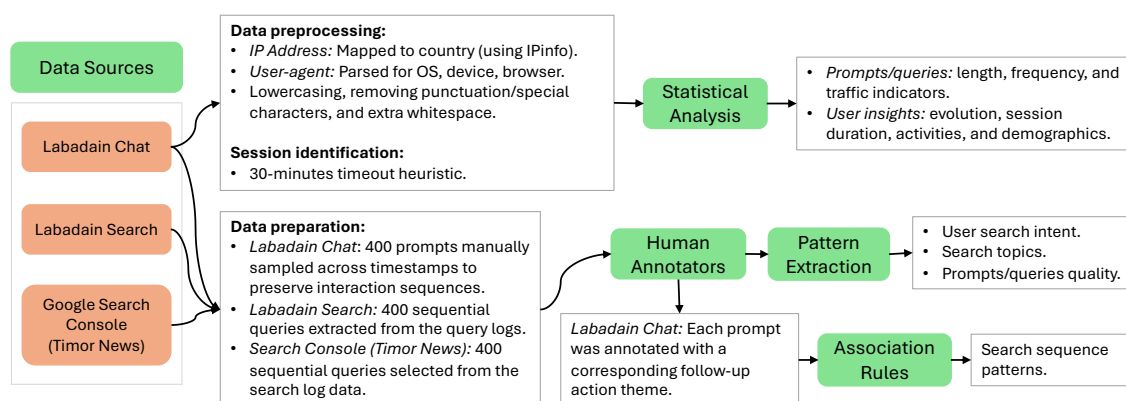


Figure 8.2: Overview of the Methodology.

To standardize the log data for analysis, prompts and queries were preprocessed by lowercasing text and removing extra whitespace, punctuation, and special characters. For Labadain Chat prompts, additional preprocessing was applied to user IP addresses and user-agent strings. Country-level geolocation was inferred from IP addresses using the free tier of IPinfo,⁸ while user-agent strings were parsed to extract information on operating system, device type, and browser.

To account for users who may have multiple visits within a single day, we segmented activity into sessions by calculating the time interval between consecutive prompts from the same user. We adopted the time-oriented heuristic proposed by Cooley et al. [46], which defines a session timeout threshold between 10 and 30 minutes. Given that the 30-minute threshold is widely used in studies of web access logs to uncover hidden patterns and navigation sequences [84, 87, 213, 248], we applied this threshold in our study.

⁸<https://ipinfo.io>

To investigate trends in user search behavior and intent on traditional search engines influenced by LLM-based conversational assistants, we compared prompt logs from Labadain Chat with query logs from Labadain Search and Google Search Console for Timor News. A total of 400 prompts/queries were sampled from each platform and distributed to four native Tetun-speaking annotators: one Ph.D. student and three third-year undergraduate students. Each annotator reviewed 300 prompts, 100 from each platform, resulting in a total of 1,200 annotated prompts.

For Labadain Chat, interaction sequences were preserved to maintain session structure and enable the analysis of follow-up prompts. Prompts were selected from four specific dates: July 22, 2024, August 1, 2024, August 31, 2024, and October 3, 2024. On each date, log collection began at 08:00, aligning with the start of increased prompt traffic. For Labadain Search and Google Search Console, 400 queries were selected sequentially in the order they appeared in the logs.

Annotators followed established guidelines to evaluate user intent, search topics, and prompt or query quality. User intent was categorized as interrogative, imperative, modal, or pronoun-based prompts [218]. Prompts/queries were further classified into three categories: information-seeking, conversational, and task-oriented. For search topic classification, we adapted the query topic categorization frameworks of Beitzel et al. [23] and Rohatgi et al. [190].

Prompt and query quality assessment focused on lexical aspects (vocabulary), grammatical accuracy, and overall clarity, regardless of any observed issues. To ensure accurate use of Tetun and its grammatical structure, annotators referenced two grammar books [78, 116] and a dictionary [49] published by the National Institute of Linguistics, which is responsible for developing the official Tetun standard.

Furthermore, to examine the sequence of user search interactions in conversational scenarios, annotators focused on Labadain Chat prompts, as the other two logs lacked sufficient follow-up data. These prompts were classified into follow-up query action themes, as described by Kim et al. [130]. Duplicated prompts were removed, and users with fewer than three interactions were filtered out to ensure that only those with at least two follow-up actions were included in the analysis. The usernames and their corresponding interaction themes/classes were then treated as basket data for analysis. To identify patterns in user search sequences, we applied association rule mining [3, 4] to automatically extract sequential search patterns from the interaction logs.

8.5 Prompt Log Analysis

The following subsections outline the key characteristics of the prompts, including their length, frequency of terms, and traffic.

8.5.1 Prompt Length

Prompts ranged from one to 1,960 words, with an average length of 9.17 words and a standard deviation of 36.93. Among the 16,952 prompts analyzed, 173 (1.02%) were single-word prompts, with “*diak*” (“good”) and “*halo*” (“make”) being the most frequent prompts. The most common prompt length was three words, accounting for 16.43% (2,785 prompts) of the total. Of these,

53.93% began with “*saida*” (“what”), contributed by 401 users (44.35%). This suggests a tendency for users to ask direct questions or make concise requests.

To further analyze prompt lengths, we categorized them into three intervals: (1) long prompts (over 100 words), (2) medium prompts (11–100 words), and (3) short prompts (10 words or fewer). Overall, 0.63% of prompts were long, 16.41% were medium, and 82.86% were short. Additionally, 1,689 prompts (9.96%) were duplicates. Detailed prompt length distributions and the total number of contributing users for each category are presented in [Table 8.2](#).

In the long prompt category, 50.93% of the prompts requested improvements in Tetun orthography, 20.37% asked for resume generation, and 16.78% consisted of text without specific instruction. The remaining prompts included requests for explanations, generating questions or examples, and translating from Tetun to English or Portuguese. The longest prompt, containing 1,960 words, was a request for a resume. This indicates users’ demand for addressing complex language-related information needs.

Table 8.2: Prompt length characteristics.

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

For medium and short prompt categories, since the total number of prompts exceeded 1,000, the “quality at a glance” technique [134] was applied by randomly selecting 100 prompts from each category to assess user intent. The analysis revealed that medium-length prompts were dominated by command-based instructions, with 12 to 19 instances of directives such as “what”, “give”, “explain”, “how”, and prompts without specific instructions. Short prompts exhibited similar patterns, with 8 to 21 occurrences of directives, including “what”, text without specific instructions, “how”, and “explain” being the most frequent directives. This trend highlights that in shorter and medium-length interactions, users predominantly rely on directive instructions to satisfy specific information needs.

8.5.2 Prompt Frequency

The analysis of prompts term frequency revealed that the most frequent unigram words were “*mak*” (3.57%, a “FOCUS marker”), “*saida*” (3.11%, “what”), and “*iha*” (2.95%, “in/at”). The word “*mak*” can also function as a “to be” verb, particularly for the third-person singular, where it is translated as “is”. For instance, the terms “*saida mak*” translate to “what is”.

Notably, “*saida mak*” was the most frequent bigram word (2.38%), with other variants of the same meaning, such as “*mak saida*” (0.28%) in 4th position, “*saida maka*” (0.20%) in 7th, and

“*sáida mak*” (0.18%) in 9th. The word “*timor leste*” (“Timor-Leste”) ranked second, “*fo tok*” (“give”) ranked 3rd, and “*iha timor*” (“in timor”) ranked 5th (see Figure 8.3).

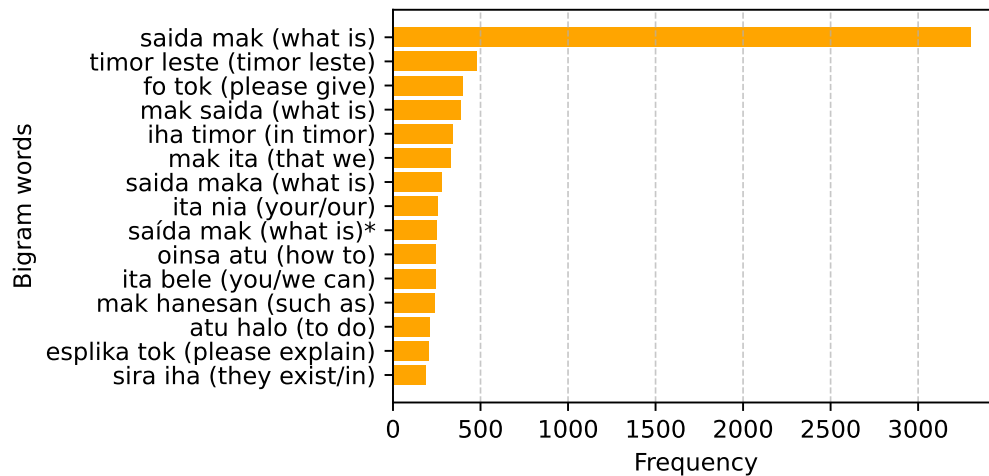


Figure 8.3: Top 15 most frequent bigrams in user prompts.*The term “*sáida*” is a misspelling caused by autocorrection, as it is similar to the Portuguese word.

Although “*iha*” ranks 3rd overall among unigrams and is primarily used to indicate place, bigrams such as “*iha timor*” (“in timor”) in 5th place and “*timor leste*” in 2nd place suggest strong search intent associated with Timor-Leste. Furthermore, user search intents such as “*oinsa atu*” (“how to”) and “*esplika tok*” (“explain”) ranked highly, in 10th and 14th positions, respectively.

In the trigram analysis, the term “*iha timor leste*” (“in Timor-Leste”) emerged as the most frequent (0.21%), suggesting that most of the requests were specific to the context of Timor-Leste. This was followed by “*saida mak ita*” (“what are you”), and “*tamba saida mak*” (“why is it”). Additionally, more specific requests appeared, such as “*esplika klaru konaba*” (“explain clearly about”) and “*hadia ortografia Tetun*” (“improve Tetun orthography”), each occurring in fewer than 100 prompts. The frequent occurrence of these bigrams and trigrams underscores users’ need for detailed explanations.

8.5.3 Prompt Traffic

To uncover insights into prompt traffic patterns, we analyzed how traffic volume fluctuates throughout the day. Figure 8.4 shows the average percentage of distinct prompts per hour over seven days. The lowest traffic occurs during the early morning hours (03:00 to 05:00), with a gradual increase starting around 06:00. Traffic peaks between 09:00 and 11:00, with a secondary peak at 21:00. The highest traffic occurs at 21:00 (7.99%), followed by 10:00 (7.65%), indicating that users are most active during mid-morning (typical working hours) and late evening.

Moreover, we analyze the variation in prompt traffic on different days of the week. The highest traffic was recorded on Wednesday (16.78%), closely followed by Monday (16.78%) and Tuesday (16.69%). Traffic decreased slightly on Thursday (16.03%) and dropped more significantly on Friday (12.22%), Sunday (11.77%), and Saturday (9.73%), with Saturday registering the lowest

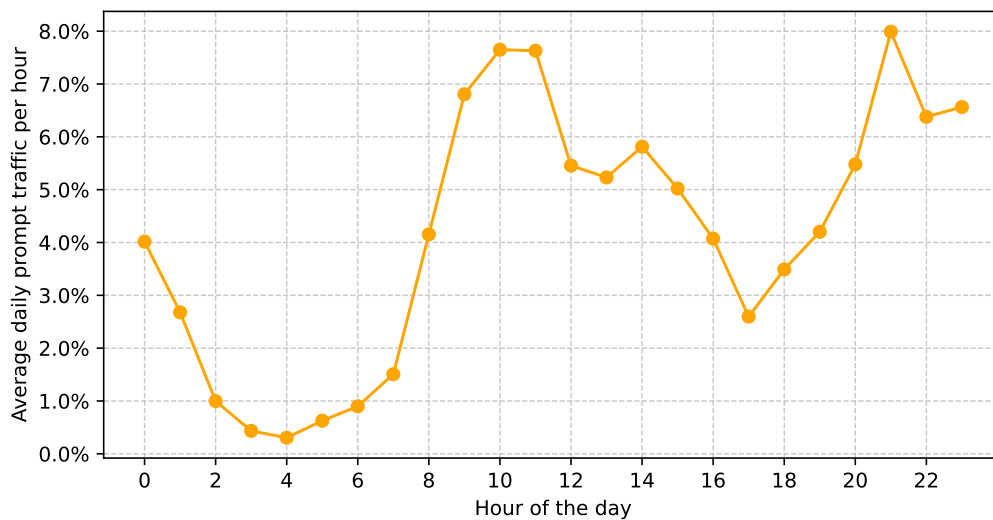


Figure 8.4: Average distribution of total and unique prompts throughout the day.

traffic overall. To gain insights into overall prompt traffic, we analyzed the average number of unique prompts per month, day, and hour. Since the records for July and October 2024 are incomplete, only August and September 2024 data were included. The analysis showed that Labadain Chat recorded an average of 6,075.5 unique prompts per month, corresponding to 201.57 unique prompts per day and 8.52 unique prompts per hour.

8.6 User Log Analysis

This section provides a detailed overview of user evolution and engagement with the platform, focusing on key metrics such as the growth of new users, returning user trends, user sessions, time spent on the platform, and demographic insights.

8.6.1 User Evolution

Labadain Chat experienced a 192.16% growth in users from the end of July to early October 2024. On average, the platform recorded 12 new users and 31 returning users per day, with weekly averages of 82 new users and 130 returning users. A detailed illustration of daily user growth and engagement with the platform is shown in [Figure 8.5](#).

There was a significant rise in new users on July 28 and 29, 2024, reaching a peak of over 80 unique users. This increase could be attributed to the introduction of improved Tetun writing features or sharing activities on social networks by registered users. During this specific period, the number of returning users remains lower than that of new users but gradually increases and maintains higher levels as time progresses, particularly from August onward.

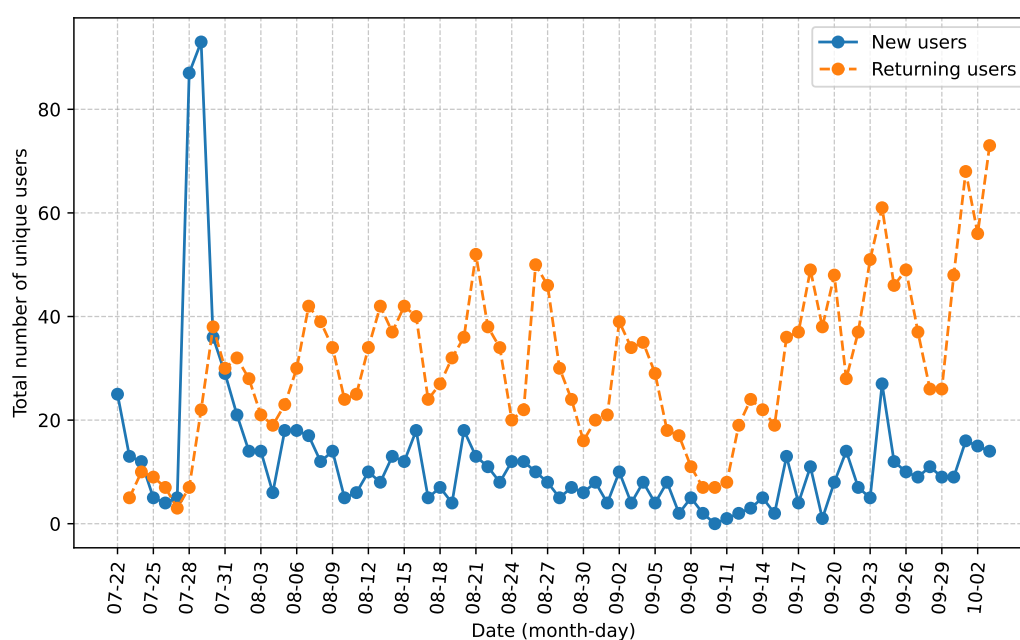


Figure 8.5: Total of unique new and returning users per day.

Returning users exhibit a more consistent upward trend, suggesting increased user retention or engagement, potentially driven by features, content generated, or services that encourage repeated use. The increase in user activity closer to October may be linked to ongoing platform improvements or external factors, such as its use for academic projects.

8.6.2 User Sessions and Activities

The analysis of user prompts showed that 99.34% of users engaged in more than one prompt, with an average of 19 prompts per user and an average per-session duration of 16.9 minutes. Moreover, 50.22% of users made more than one session, with an average of 4.11 sessions per user.

During each session, 93.97% of users engaged in multiple prompts, with an average of 3.89 prompts per session and an average session duration of 16.9 minutes. In terms of sessions over time, the platform recorded an average of 229 sessions per day from 43 unique users and 1,541 sessions per week from 181 unique users. Detailed metrics on user prompts and sessions are provided in [Table 8.3](#).

Table 8.3: Summary of user prompts and sessions.

Metric	Min.	Max.	Avg.	Std.
Prompts per user	1.00	310.00	18.75	29.18
Prompts per session	1.00	53.00	3.89	4.07
Session per user	1.00	42.00	4.11	5.17
Time spent per session (min)	0.23	330.82	16.9	20.59

Based on the highlighted data in Table 8.3, we conducted a detailed analysis of the most active user, who engaged in a total of 310 prompts (highlighted in orange) and spent a total of 917 minutes (15.28 hours) on the platform. The longest time spent in a single session was 331 minutes (5.5 hours, highlighted in green), consisting of 53 prompts, with an average time spent per prompt of 6.24 minutes (min. 1.65 minutes, max. 27 minutes, std. 4.9). These findings indicate that certain users demonstrate high engagement, either through frequent prompts over time or extended sessions in one sitting.

Figure 8.6 provides further insights into user activity patterns throughout the day, revealing that new users exhibit a distinct spike in activity during the early morning, followed by a more dispersed pattern over the remaining hours. It focuses on the total number of users and unique users within the same 24-hour period. Similar to Figure 8.4, it shows peaks in total user activity during the morning and evening hours, largely driven by returning users. In contrast, the number of unique users remains relatively stable throughout the day, suggesting that a consistent group of users accounts for most of the activity.

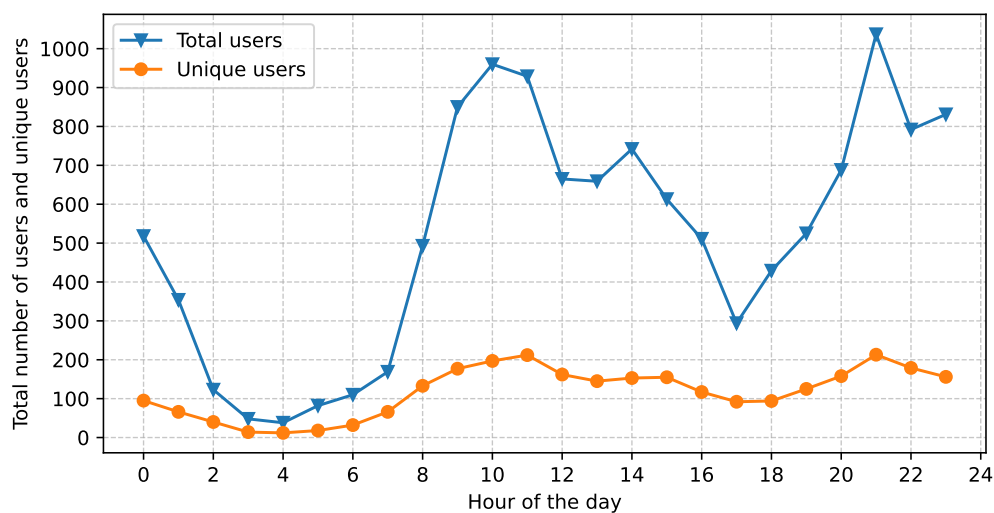


Figure 8.6: Total number of users and unique users per hour.

In terms of time spent during each session, most users spent between one and five minutes, followed by 16 to 30 minutes, and six to ten minutes. This distribution, illustrated in Figure 8.7, highlights the variability in user engagement during their visits.

8.6.3 User Demographics

User demographics were analyzed using two data sources: self-reported information provided during user registration and data extracted from IP addresses and user agents in the prompt logs.

8.6.3.1 Self-Reported

According to self-reported data, most of the users (74.78%) are aged 18 to 25, followed by 20.58% aged 26 to 40. A smaller proportion (2.43%) is under 17, while 2.21% are over 40. In terms of gen-

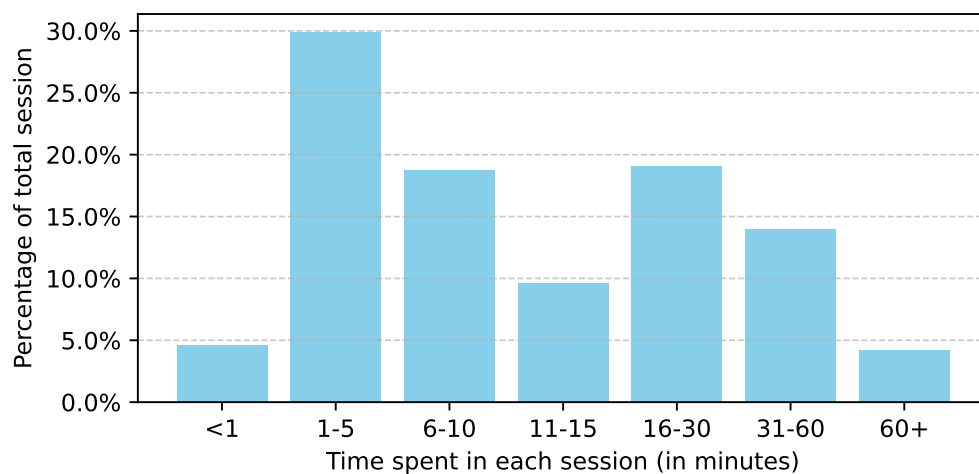


Figure 8.7: Time spent by users per session.

der distribution, 62.83% of users are male and 37.17% are female. Almost all users (99.56%) are located in Timor-Leste, spread across various municipalities, with the majority (65.71%) residing in Dili, the capital of Timor-Leste.

8.6.3.2 User Log Analysis

The analysis of user log data provides insights into user locations, which are identified through IP addresses, as well as the platforms and devices used to access the system, derived from the user agent information.

User Locations: Consistent with self-reported data, most of the users (95.22%) are located in Timor-Leste, followed by Singapore (2.11%) and Indonesia (1.02%). Other locations, including Australia, Portugal, South Korea, Brazil, and the United States, among others, each account for less than 1% of the total.

Operating Systems: Android is the most widely used operating system, accounting for 70.88% of users, followed by Windows (17.23%) and iOS (8.69%), with all other operating systems contributing less than 2%.

Browsers: Most of the users (88.97%) access the platform through Google Chrome, with smaller shares using Safari (8.94%) and Firefox (2.03%).

Device Types: Smartphones are the predominant device type, used by 79.54% of users. Desktops and laptops account for 20.34%, while tablets and other devices represent less than 1%.

8.7 Traditional and LLM-Based Search

This section presents an analysis of log data from two conventional search platforms and one LLM-based conversational assistant. The analysis includes a comparison of prompts manually annotated by native Tetun speakers, focusing on key aspects such as user search intents, search topics and sequences, and prompt/query quality.

8.7.1 User Search Intent

The user intents were categorized following the definitions of Trippas et al. [218], classifying prompts into interrogative, imperative, modal, and pronoun-based categories. In Figure 8.8, the blue bar represents conversational search interactions on Labadain Chat, the green bar corresponds to searches performed using the Labadain Search engine, and the red bar represents searches conducted via Google and captured by the Google Search Console (labeled as “Timor News”).

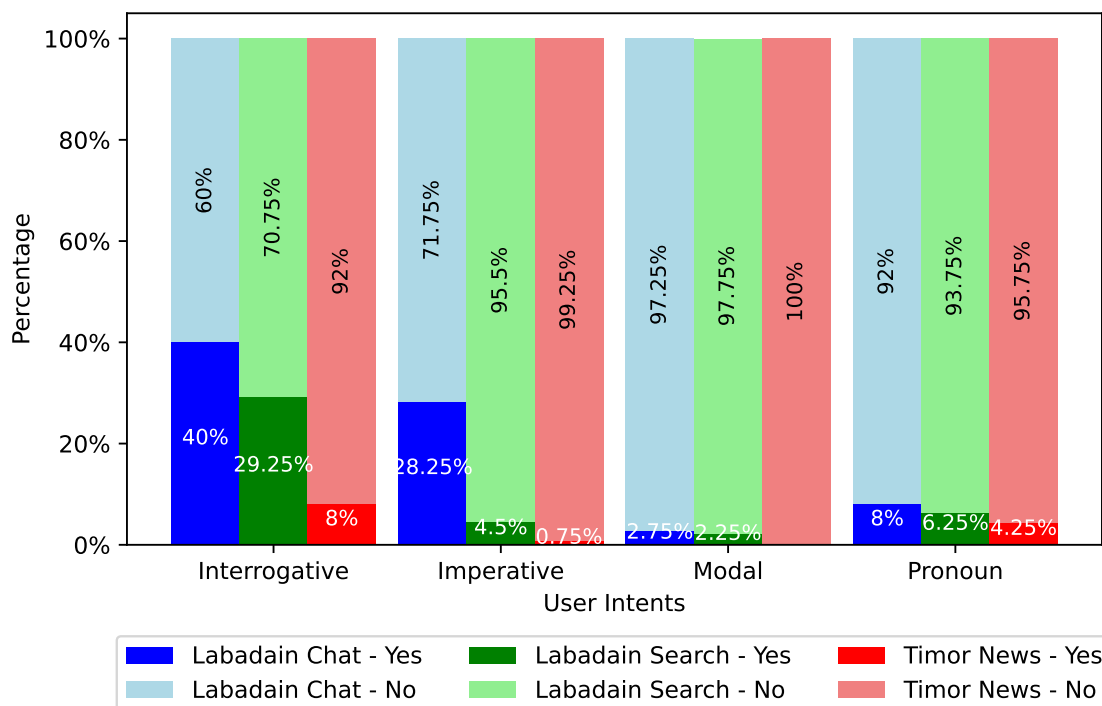


Figure 8.8: User intent from prompt analysis by human annotators across three platforms: Labadain Chat (conversational search), Labadain Search (keyword search), and Timor News (keyword search from Google). The “yes” indicates the presence of the respective class in the prompt (i.e., interrogative, imperative, modal, or pronoun), while “no” indicates its absence.

Overall, Figure 8.8 shows small differences (less than 10%) in the use of modals and pronouns in user search queries on the traditional Google search engine and Labadain Search and the LLM-based search on Labadain Chat (represented by the two bar groups on the right). In contrast, substantial differences (over 20%) are observed for imperative and interrogative intents. These findings emphasize the prevalence of command-based instructions in Labadain Chat.

In terms of search intent distribution, the prompts were categorized into four main categories [120, 181, 182]: *Information-seeking* (queries aimed at obtaining factual information, knowledge, or specific details about a topic), *Task-oriented* (queries where users instruct the system to perform a specific action or complete a particular task), *Conversational* (interactions that mimic human-like dialogue where the user engages in a casual conversation with the system), and *Other* (miscellaneous prompts, including queries to test the system's capability). The analysis results for Labadain Chat, Labadain Search, and Timor News reveal distinct patterns in the search intents across the three platforms, as presented in Table 8.4.

Table 8.4: Comparison of search intent across conventional and LLM-based platforms.

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

Information-seeking queries were dominant on all platforms, accounting for 87.75% of prompts in Labadain Chat, 98.00% in Labadain Search, and 99.50% in Timor News. Task-oriented prompts were only observed in Labadain Chat (10.50%), indicating its role as a conversational assistant capable of performing specific tasks, while no such queries were recorded on Labadain Search or Timor News. Conversational interactions were minimal, with Labadain Chat showing 1.75% and Labadain Search 2.00%, while no such queries were observed in Timor News.

The analysis in Figure 8.8 and Table 8.4 reveals a trend in the user search behavior on Labadain Search, a conventional search engine developed after the emergence of LLM-based systems. In the leftmost group of bars in Figure 8.8, a higher frequency of interrogative-based commands is observed in Labadain Search compared to search queries captured by the Google Search Console for Timor News, showing a difference of +21.25 percentage points.

This finding is supported by the search intents shown in Table 8.4, where 1.50% of the queries in Labadain Search include conversational elements such as “*ita diak ka lae*” (how are you?) and “*haii boatardi*” (hi, good afternoon). These observations suggest that LLM-based systems have influenced traditional search behavior, underscoring the need for search system solutions that effectively adapt to evolving user behaviors driven by LLM-based conversational assistants.

8.7.2 Search Topics and Sequences

To classify search topics, we analyzed query topics reported by Beitzel et al. [23] and Rohatgi et al. [190] to define topic categories relevant to this study. The process resulted in 17 topics, each with detailed descriptions provided in Table A.5 of Appendix A. The annotation results are presented in Table 8.5, highlighting key differences in search topic distribution across the platforms.

Labadain Chat recorded the highest proportion of prompts related to education/science, health, social/culture, and research/academia, followed by Labadain Search. In contrast, Labadain Chat

Table 8.5: Comparison of search topics between conventional and LLM-based platforms.

Topic	L. Chat	L. Search	Timor News
Education/Science	28.75%	16.75%	11.25%
Health	28.00%	16.50%	12.25%
Social/Culture	13.75%	11.75%	3.25%
Research/Academia	6.75%	4.00%	3.75%
Politic/Government	2.50%	7.50%	6.00%
Job vacancy	0.00%	3.50%	7.25%
Economy/Finance	1.00%	6.75%	2.75%
Agriculture/Environment	6.50%	2.00%	1.75%
Law/Justice	0.75%	3.75%	1.75%
Language/Translation	2.00%	4.25%	1.50%
Computer science	3.75%	3.00%	1.25%
Personal	0.50%	2.50%	3.25%
Porn	0.00%	0.25%	2.00%
Sport	0.75%	1.00%	0.50%
Holiday/Travel	0.25%	0.50%	0.75%
Entertainment	0.00%	0.50%	0.75%
Other	4.75%	15.75%	40.00%

showed the lowest percentage of prompts for politics/government and economy/finance, which were more prevalent in Labadain Search, with 7.50% and 6.75%, respectively.

Timor News exhibited a distinct focus, with the highest percentage of queries related to job vacancy (7.25%), compared to Labadain Search at 3.50%, and no such queries were recorded in Labadain Chat. Additionally, Timor News had the largest proportion of queries categorized as “Other” (40.00%), attributed to the difficulty of categorizing short queries, since 75.75% of the total queries consisted of just one to three words [e.g., “*atribui*” (assign), “*karta pedidu*” (“request letter”), “*4 de setembru*” (4 September)].

In the analysis of follow-up search sequences, 400 prompts from 91 users were initially annotated by human annotators into 12 themes based on the actions of the follow-up queries described by Kim et al. [130]. Among these, 17 repeated prompts and interactions from 58 users with fewer than three occurrences were excluded. Consequently, 237 interactions (representing 59.25% of the annotated data) of 33 users were included in the sequence pattern analysis. Sequence patterns were calculated using a minimum support of 0.1, a minimum confidence threshold of 0.5, and a minimum lift of 1. The top five sequence search rules are presented in Figure 8.9.

Figure 8.9 represents the analysis of the follow-up search sequences, showing how users refine their queries after initial searches. The strongest sequence, highlighted by the red arrow with a lift of 1.31 (support 0.31, confidence 0.83), indicates that users frequently transition from substituting a condition to adding/specifying a condition. Other observed sequences, with lift values ranging from 1.10 (support 0.12, confidence 0.5) to 1.28 (support 0.21, confidence 0.58), reflect additional patterns, such as users excluding conditions or requesting related information before further adjustments. This visualization reveals that users often engage in iterative query refinement to

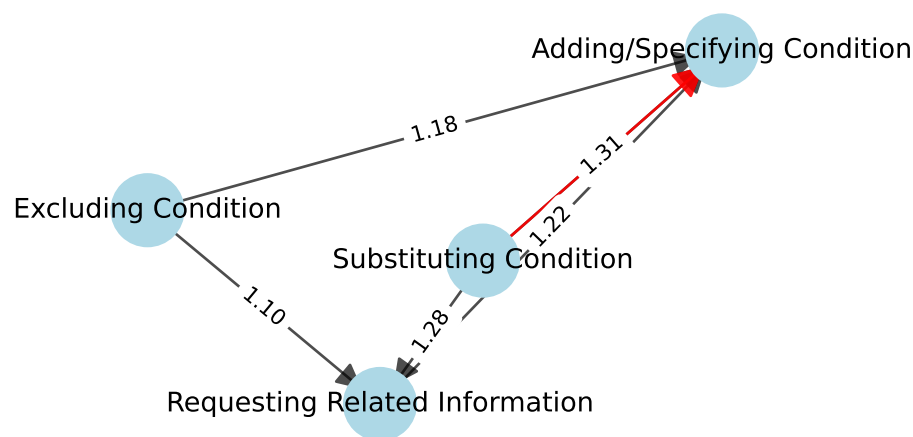


Figure 8.9: Visualization of association rules in user follow-up search sequences. Edges represent relationships between themes, with arrow directions indicating the sequence of actions. Edge labels denote lift values, and the red arrow highlights the strongest association with a lift of 1.31.

narrow their search results or clarify their intent.

8.7.3 Prompt/Query Quality Evaluation

The evaluation of prompt and query quality considered three dimensions: lexical quality, which assessed word-level correctness; grammatical quality, which examined word correctness and grammatical structure; and overall quality, which considered clarity and specificity regardless of lexical or grammatical errors. For each dimension, we applied a five-point scale ranging from excellent (indicating no errors) to unacceptable (denoting severe impairments to understanding), with intermediate ratings of good, fair, and poor reflecting varying degrees of errors.

In terms of lexical quality, Timor News achieved the highest rating, with 65.5% of prompts rated as excellent, followed by Labadain Search at 41.25% and Labadain Chat at 18.75%. The lower rating for Labadain Chat is attributed to its longer prompts, which introduced more lexical errors than the shorter queries in traditional keyword searches. A similar trend was observed in grammatical quality, where Timor News again led with 27% rated as excellent, followed by Labadain Search at 12.75% and Labadain Chat at 10.5%.

For overall quality, Labadain Chat achieved the highest rating, with 29% of the prompts rated as excellent, compared to 8.25% for Timor News and 7.75% for Labadain Search. This higher score reflects the specificity and structured nature of Labadain Chat prompts, which contributed to greater clarity and understanding, offering an advantage in the overall quality assessment.

8.8 LabadainLog-17k+ Dataset Description

LabadainLog-17k+ [71] is a dataset comprising log data collected from three platforms: Labadain Chat, Labadain Search, and Google Search Console for Timor News. The dataset contains 16,952 conversational prompts from Labadain Chat, 400 queries from Labadain Search, and 400 search

queries from Google Search Console for Timor News. All data were collected during the same period, from July 22 to October 3, 2024.

Table 8.6: Examples of log data entries in LabadainLog-17k+.

Labadain Chat	
2024-07-27 08:35:48	Timor-Leste Mozilla/5.0 (Linux;... G8rXbFZ5 Saida Mak urinaria Saida mak urinária?
Labadain Search	
wxqlicbn747aodx1o2v1pv0r8h4224he	Timor-Leste 2024-10-03 08:22:26.459380 Saida mk ema Saida mak ema?
Timor News	
kazu moras dengue	Kazu moras denge

The Labadain Chat logs include request timestamps, user IP addresses, user agents, user-names, and prompts. The Labadain Search logs comprise user session IDs, IP addresses, request timestamps, and queries, while the Google Search Console logs only contain search query text. Table 8.6 provides examples of LabadainLog-17k+ log entries, with each entry separated by a newline, and the elements within each entry are delimited by a tab.

To protect user privacy [225], all personally identifiable information has been anonymized, while preserving the sequence of interactions to retain the contextual information about user behavior. Usernames in the Labadain Chat log data were anonymized (highlighted in orange in Table 8.6) using the secure one-way hashes technique [47], and user IP addresses were resolved to country names (highlighted in green in Table 8.6) using IPinfo. The secure one-way hash method replaces each username with a unique hash value while ensuring consistency across the dataset. The same username is consistently mapped to the same hash. To improve usability and maintain concise identifiers, the hash length was reduced to six bytes while still preserving both uniqueness and security.

The dataset included country and user-agent information to enable the replication of this study. User-agent data, which captures information about the device and browser used, is included to enhance the understanding of platform accessibility. Additionally, prompts and queries with corrected orthography are provided alongside the original versions (highlighted in red in Table 8.6) to enable broader applicability. Explicit user consent was obtained before collecting user data, ensuring ethical compliance and user privacy.

8.9 Results and Discussion

The analysis of prompt logs revealed that 82.86% of the prompts were short, containing between one and ten words. Among all prompts, three-word prompts were the most common, accounting for 16.43% of the total. Of these, 53.93% began with “what”, contributed by 44.35% of users. Additionally, unigram, bigram, and trigram word prompts were dominated by function words, “timor”, “timor leste”, and command-based words such as “what”, “how”, and “explain”.

Platform activity peaked during mid-morning and late evening, with returning users driving engagement throughout the day. Traffic was particularly high at the beginning and middle of the week, gradually decreasing toward the weekend. Over half of the users made multiple sessions, spending an average of 16.9 minutes per session. During each session, most users spent between one and five minutes, averaging 3.89 sessions. Most of the users were young men aged 18 to 25, predominantly located in Timor-Leste, with a strong preference for accessing the platform via mobile devices and the Google Chrome browser.

Responding to the RQ1, comparing the findings of this study with those of Trippas et al. [218], several similarities and differences emerge, as summarized in Table 8.7. Both studies reveal similar engagement metrics, including the frequent use of action-oriented instructions (e.g., “explain” and “give”), which emphasize direct and task-driven requests, and the dominance of function words (e.g., that, in/at). Additionally, both studies observed a small proportion of single-prompt interactions (1.02% vs. 1.70%), reflecting a tendency toward brevity in queries. The average session duration (16.9 minutes in this study vs. 4 minutes in Trippas et al. [218]), indicating comparable user interaction patterns across platforms despite differences in design and target audience. Furthermore, the longest session durations (330.82 minutes vs. 194 minutes) suggest that certain users exhibit prolonged engagement with the platforms, highlighting their utility for specific tasks.

Table 8.7: Key similarities and differences between this study and Trippas et al. [218].

Aspect	This study	Trippas et al.
Similarities		
Total of Prompts Analyzed	16,952	14,595
Total Unique Users	904	95
Avg. Sessions per User	19	49
Avg. Time Spent per Session	16.9 mins	4 mins
Longest Session Duration	330.82 mins	194 mins
Proportion of Single Prompts	1.02%	1.70%
Avg. Prompt Length	9.17 words	43.35 words
Frequency of Function Words	Dominant	Dominant
Action-Oriented Words	Frequent	Frequent
Differences		
Prompt Language	Tetun	English
Data Source	Labadain Chat	Google Bard
Data Collection Method	System	Crowd Workers
User Search Experience	No Data	Self-Reported
User Feedback	No Data	Self-Reported
User Search Topics	Analyzed	Not Analyzed
Follow-Up Search Sequence	Analyzed	Not Analyzed
User Evolution	Analyzed	Not Analyzed
Temporal User Engagement	Analyzed	Not Analyzed

The differences between the two studies are pronounced, reflecting distinct languages, data

characteristics, collection methods, research objectives, user demographics, and contextual settings. This study focuses on Tetun, a LRL that offers insights into search patterns among Tetun-speaking users. In contrast, Trippas et al. [218] examine English, a highly resourced language, with users spanning six countries across three continents, providing a global perspective.

The data sources and collection methods also differ: this study analyzes log data collected through real-world user search activities, whereas Trippas et al. [218] examine prompts from crowd workers and self-reported feedback. Notably, the nature of real-world user data often diverges from self-reported feedback, as it reflects actual user behavior rather than perceived or intended actions. For example, users may report that they primarily search for work-related information, while their actual behavior often includes entertainment- or curiosity-driven prompts.

Furthermore, this study explored search topics, user evolution (new vs. returning), temporal engagement, follow-up query sequences, and query categorization—areas not covered by Trippas et al. [218]. Meanwhile, Trippas et al. [218] emphasized user feedback and search experience, incorporating subjective insights into satisfaction and interaction quality, while this study relied entirely on objective system data. These findings highlight the importance of studying user search behavior across platforms, languages, and contexts to design search systems that accommodate both universal patterns and context-specific user needs.

In response to RQ2, this study observed a trend toward the use of command-based instructions in user search behavior on conventional search engines (see Figure 8.8), suggesting that LLM-based systems have influenced search behavior on traditional platforms by aligning queries more closely with the typical conversational prompts used in LLM-based assistants. This finding underscores the need for search systems to effectively adapt to evolving user behaviors shaped by the rise of LLM-based conversational systems.

Regarding search topics, distinct priorities emerged across platforms: Labadain Chat users primarily focused on education/science and health, Timor News users emphasized job vacancies, and Labadain Search users concentrated on politics/government and economy/finance. In follow-up search sequences, users iteratively refined their prompts after initial searches to narrow results or clarify intent. Prompts from Labadain Chat were generally more specific and structured, but their longer and more complex forms showed higher lexical and grammatical errors.

Moreover, comparing Labadain Chat interaction logs with those of two traditional search engines shows distinct search topics, a trend in query formulation within traditional search engines, and a higher occurrence of lexical and grammatical errors in queries submitted to LLM-based assistants. These errors are likely influenced by Timor-Leste’s multilingual landscape, users’ proficiency in official Tetun, and limited experience with formal query construction.

8.10 Limitations and Potential Biases

While this study provides valuable insights into conversational search behavior and identifies common patterns shared with the findings of Trippas et al. [218], we acknowledge several limitations and potential biases. First, the dataset focuses exclusively on Tetun-speaking users, which, while

essential for studying an LRL, may limit the generalizability of the findings to broader linguistic and cultural contexts. Additionally, due to the short data collection periods, longer-term evolutions in user behavior may not be fully captured. Potential bias may arise from the nature of real-world user log data. While such data provides authentic insights, the absence of controlled experimental conditions makes it challenging to associate behaviors with specific user profiles and to determine the influence of LLM-based systems on search trends in conventional platforms among the same user groups. Since Labadain Chat users are required to register and are aware of the data collection for research purposes, their behavior may have been influenced by perceived expectations.

8.11 Conclusions

This study investigates user search behaviors, engagement patterns, and intent through interaction logs from a real-world LLM-based conversational system designed for the Tetun-speaking community. By analyzing search intents, topic distribution, and prompt quality, we compare these behaviors with data from two conventional search engines serving the same linguistic group. Additionally, our findings are contextualized through a comparative analysis with English-speaking user interactions reported by Trippas et al. [218].

The findings from this study reveal that over half of registered users (50.22%) engaged in multiple sessions, with an average of 3.89 prompts per session and an average session duration of 16.9 minutes. The steady rise in new user registrations suggests increasing user retention and sustained engagement (see [Figure 8.5](#)), potentially influenced by platform features, ongoing improvements, and the quality of generated responses.

Furthermore, the prevalence of search topics related to education, science, and health underscores the perceived value of Labadain Chat in addressing user information needs. While further research is necessary to fully assess the system's impact on the Tetun-speaking community and its alignment with their specific requirements, user engagement patterns provide strong evidence of its relevance and utility.

The comparative analysis of Tetun-speaking user search behaviors across traditional and LLM-based platforms reveals a notable trend in query formulation within the newer search engine, where queries increasingly resemble prompts typical of LLM-based assistants. This trend suggests that LLM-driven technological advancements are influencing user search behaviors, even among Tetun-speaking users.

By examining the context-specific nature of user search behaviors across platforms, regions, and languages, this study uncovers notable parallels with the findings of Trippas et al. [218], including similarities in the predominance of short prompts, the frequent use of action-oriented instructions, and the dominance of function words. These similarities are particularly noteworthy given the substantial differences between the two studies—including system architecture, interaction log data, language (Tetun versus English), and user demographics—suggesting that certain user behaviors may persist across diverse technological and linguistic contexts.

At the same time, key differences emerge in data sources, collection methodologies, dataset characteristics, and user attributes analyzed. These findings address the research questions by offering meaningful insights into the evolving nature of user search behaviors across linguistic contexts and underscore the importance of search systems to dynamically adapt to these trends.

To support further research on LLM-based search behaviors in low-resource languages, we publicly release LabadainLog-17k+ [71], under the Creative Commons Attribution Share-Alike (CC BY-SA) license. This dataset offers a unique resource for analyzing conversational AI interactions and tracking the evolution of search behavior within the Tetun-speaking community.

Chapter 9

Conclusions

This work represents a pioneering effort in the development of information retrieval (IR) for Tetun, establishing a foundation for future advancements in Tetun language technology through the development of essential language processing tools and resources. Beyond its specific contributions, it introduces methodologies that contribute to broader discussions in addressing language development challenges for low-resource languages (LRLs). The following sections summarize the key contributions, findings, limitations, open issues, and future directions of this work.

9.1 Contributions to Tetun IR and NLP

This work has made significant contributions to the development of foundational linguistic tools and resources for Tetun, which were previously unavailable. It introduced the first tokenizer [67] capable of effectively segmenting Tetun text, designed with careful consideration of the language’s unique morphological characteristics. Additionally, a language identification model [66] was developed to distinguish Tetun from other languages commonly spoken in Timor-Leste. These tools establish a crucial foundation for core language processing tasks, enabling effective digital content processing in the country’s multilingual context.

A rule-based stemmer [75] was designed to handle Tetun inflections and derivations, while also addressing the challenges posed by Portuguese loanwords to ensure accurate word normalization. A stopword list [72] was compiled to facilitate IR and natural language processing (NLP) research. Furthermore, a text dataset enriched with metadata [64] was developed, offering a unique resource to enable future research in the IR and NLP domains.

To advance IR research in Tetun, a test collection [70] and baselines [73, 77] for ad-hoc text retrieval were developed to evaluate system performance. Baselines for traditional sparse retrieval methods were established [73], incorporating language-specific components to serve as a strong reference point for future work. Additionally, neural-based retrieval approaches were explored and a baseline was established [77], comprising results from zero-shot and hybrid retrieval strategies that leverage large language models (LLMs) to enrich document representation, which leads to

improved retrieval performance. All run files and document summaries [76] associated with these experiments have been made publicly available to enable reproducibility.

A search log dataset [71] capturing real-world user interactions in Tetun, sourced from traditional search engines and an LLM-based conversational assistant, was also released to the public. This resource offers a unique opportunity for further research into the search behavior and information access patterns of Tetun-speaking users. Collectively, these contributions provide a robust foundation for future research and development in Tetun IR and language processing.

By introducing these foundational IR and NLP resources and tools for Tetun, this work enhances the language’s digital presence, supporting its integration into the global digital ecosystem and enabling sustained progress in IR and NLP research. The development of these linguistic tools and resources not only contributes to language preservation but also promotes digital inclusion and equity, ensuring that the Tetun-speaking community can fully participate in the global digital age.

9.2 Findings on Tetun Linguistic Aspects

This work has uncovered several key insights into the current state of Tetun in the digital landscape. One notable observation is that Tetun documents are predominantly sourced from online newspapers [62], with written forms increasingly adopting the standard established by the *Instituto Nacional de Linguística* (INL). This trend reflects growing consistency in orthographic conventions and suggests progress in the adoption of the official Tetun standard across digital platforms. Furthermore, the continued and prominent use of Portuguese loanwords in Tetun texts highlights the need for NLP tools to effectively manage the dynamic interplay between native vocabulary and borrowed terms.

Another critical finding is the need for continuous updates to Tetun linguistic resources. Existing dictionaries and lexicons often did not capture the evolving nature of the language, particularly with the emergence of new terms and expressions. This gap emphasizes the importance of regularly refining and expanding these resources to maintain their relevance and effectiveness in supporting the development of language technology for Tetun. These findings underscore the necessity for sustained and adaptive efforts to ensure that linguistic resources keep pace with the dynamic evolution of the language.

9.3 Contributions to Low-Resource Languages IR

Beyond the scope of Tetun, this work offers methodological contributions that can be applied to other LRLs. A key contribution is the development of a lightweight data collection pipeline [63] designed to streamline and automate web crawling tasks. This pipeline facilitates efficient corpus construction in contexts where online documents exist but lack readily accessible and structured resources. Moreover, a robust methodology for data annotation and processing [62] has been proposed, establishing a systematic approach to developing and characterizing web-crawled documents for under-resourced languages.

The work also introduces a method for stopword detection [9, 73], which can be adapted to other LRLs that lack predefined stopword lists. Additionally, the development of a rule-based stemmer [75] that accounts for loanwords provides a conceptual model for languages with similar linguistic influences, offering a framework for handling borrowed vocabulary. Another significant contribution is the detailed procedure for test collection development [73], addressing a common gap in prior research, where such methodologies were often insufficiently documented. Furthermore, the methodology for establishing ad-hoc text retrieval baselines [73] ensures that language-specific characteristics are carefully considered when evaluating IR systems for LRLs.

Finally, this work proposes a document contextualization approach [77], which enriches document titles with content summaries generated by an LLM, to support the evaluation of the performance of pretrained dense retrieval models in zero-shot and hybrid strategies. This technique provides a promising alternative in scenarios where passage-level collections are unavailable, a common limitation in many LRLs, to evaluate dense retrieval models that typically rely on richer contextual signals. Collectively, these findings offer practical methodologies and frameworks to advance IR and NLP research for LRLs.

9.4 Limitations and Open Issues

Despite the progress achieved in Tetun IR and language processing, several limitations and challenges remain. A primary concern is the scarcity of up-to-date linguistic resources for Tetun, such as dictionaries, books, academic literature, and other relevant materials. While this work has contributed resources and tools, the linguistic landscape of Tetun continues to evolve, requiring regular updates to capture contemporary language use. The integration of newly emerging words, often shaped by ongoing linguistic change, remains a persistent challenge.

Another notable limitation is the issue of standardization. Although the INL standard is widely recognized, variations in Tetun orthography and writing practices persist. This inconsistency presents challenges for the development of language technologies, underscoring the need for NLP tools that can effectively handle orthographic variation.

Furthermore, limited funding restricted the involvement of Tetun linguists in this work, narrowing the scope of linguistic validation, despite their expertise being invaluable for evaluating linguistic resources and guiding rule-based approaches. Consequently, certain linguistic aspects were not thoroughly examined as intended. Expanding collaborations to include Tetun native linguists could enhance the robustness of computational linguistic models and algorithms.

Finally, funding limitations also present a broader challenge to the sustainability of Tetun IR and NLP research. The development and maintenance of linguistic resources require continuous investment, yet research on LRLs like Tetun often receives minimal to no financial support. Bridging this funding gap is essential to ensure the long-term advancement and impact of technologies associated with the computational processing of Tetun.

9.5 Future Directions

Building on the findings and contributions of this work, several future research directions can be pursued. Expanding Tetun linguistic resources remains a key priority, particularly through the collection and annotation of larger, high-quality datasets to support continued progress in Tetun IR and NLP research. Future work can also focus on improving tools, algorithms, and techniques introduced in this work by integrating more advanced machine learning techniques and LLMs, in line with ongoing developments in the IR and NLP domains.

While retrieval baselines have been established, further research into advanced search solutions for Tetun remains an important research direction. In particular, the development of a Tetun passage-level dataset and corresponding test collection would align with current trends in text retrieval and facilitate Tetun’s integration into multilingual and cross-lingual research. Leveraging LLMs to further enhance Tetun text retrieval effectiveness also presents a promising opportunity. Furthermore, expanding user behavior analysis through the study of search log data could inform better query processing techniques and improve user experiences within Tetun search systems.

Cross-lingual information retrieval (CLIR) represents another important direction for future exploration. Given Timor-Leste’s multilingual context, CLIR techniques could enable access to information across languages, bridging linguistic divides in digital ecosystems. Finally, the investigation of multimodal retrieval for Tetun presents a valuable path toward broadening the scope of information access beyond text-only systems.

9.6 Final Remarks

This work has established a foundational framework for Tetun IR and NLP through the development of essential linguistic tools and resources tailored to the language. The findings emphasize the importance of aligning language technology development with established linguistic standards while addressing the challenges posed by borrowed vocabulary and the evolving nature of Tetun.

Beyond its contributions to Tetun, the methodologies presented in this work offer valuable insights into IR and language processing approaches for LRLs more broadly. By detailing systematic processes for data collection, annotation, characterization, and the creation of a test collection and retrieval baselines, this work provides a replicable framework that can be adapted to other linguistic contexts facing similar resource constraints.

Looking ahead, the continued development of Tetun IR and NLP tools will be critical to ensuring that the language’s digital presence keeps pace with technological advancements. Sustained research efforts, supported by adequate funding, will be essential to drive further innovation in Tetun language technology. The progress achieved in this work lays a foundation for developing more robust language processing for Tetun, while also paving the way for the digital representation and preservation of other under-resourced languages.

Appendix A

Appendices

This section provides supplementary materials and technical details supporting this study. Section A.1 outlines the Tetun Stemmer algorithms, while the criteria for assessing the quality of web-crawled documents are presented in Section A.2. Section A.3 introduces an algorithm for automated content annotation, and the Tetun stopwords list is provided in Section A.4. Finally, Section A.5 provides the descriptions for the user search topic categories.

A.1 Tetun Light Stemmer Algorithm

The linguistic regions used in the Tetun stemmer (*Labadain-Stemmer*) are shown in Table A.1, while Table A.2 lists Portuguese-derived suffixes. The algorithm for the Tetun *light* stemmer variant is detailed in Algorithm 1.

Table A.1: Linguistic regions used in the *Labadain-Stemmer*.

Region	Definition
R1	The region starting after the first non-vowel that follows a vowel, or, if no such non-vowel exists, it is the null region at the end of the word.
R2	The region starting after the first non-vowel that follows a vowel within R1, or, if no such non-vowel exists, it is the null region at the end of the word.
RV	If the second letter is a consonant, RV starts after the next vowel. If the first two letters are vowels, they begin after the following consonant. In the case of a consonant-vowel combination, RV starts after the third letter. If none of these conditions are met, RV starts at the end of the word.

Table A.2: List of Portuguese-derived suffixes.

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

Algorithm 1: Tetun Light Stemmer Algorithm**Require:** *R1*, *R2*, *RV*, *word_list***Require:** *general_suf* $\leftarrow$ list of **general** suffixes**Require:** *lojia_suf* $\leftarrow$ list of **lojia** suffixes**Require:** *usaun_suf* $\leftarrow$ list of **usaun** suffixes**Require:** *ensia_suf* $\leftarrow$ list of **ensia** suffixes**Require:** *amente_suf* $\leftarrow$ **amente** suffix**Require:** *iv_suf* $\leftarrow$ **iv** suffix**Require:** *at_suf* $\leftarrow$ **at** suffix**Require:** *ozikad_suf* $\leftarrow$ list of **ozikad** suffixes

Require: *mente_suf* $\leftarrow$ **mente** suffix

Require: *ante_suf* $\leftarrow$ list of **ante** suffixes

Require: *idade_suf* $\leftarrow$ list of **idade** suffixes

Require: *abil_suf* $\leftarrow$ list of **abil** suffixes

Require: *iva_suf* $\leftarrow$ list of **iva** suffixes

Require: *verb_suf* $\leftarrow$ list of **verb** suffixes

Require: *residual_suf* $\leftarrow$ list of **residual** suffixes

```

1: for all word in word_list do                                     ▷ Step 1: Word length validation
2:   if  $\text{length}(\text{word}) < 4$  then
3:     Return word
4:   else                                                             ▷ Step 2: Standard suffix removal
5:     if word ends with any suffix in general_suf sorted by length descending then
6:       if Position of suffix in word is in R2 then
7:         Delete suffix from word
8:         stem  $\leftarrow$  word without suffix
9:         Return stem
10:      end if
11:     else if word ends with any suffix in lojia_suf then
12:       if Position of suffix in word is in R2 then
13:         Replace suffix in word with loj
14:         stem  $\leftarrow$  word without suffix concatenates with loj
15:         Return stem
16:       end if
17:     else if word ends with any suffix in usaun_suf then
18:       if Position of suffix in word is in R2 then
19:         Replace suffix in word with u
20:         stem  $\leftarrow$  word without suffix concatenates with u
21:         Return stem
22:       end if
23:     else if word ends with any suffix in ensia_suf then
24:       if Position of suffix in word is in R2 then
25:         Replace suffix in word with ente
26:         stem  $\leftarrow$  word without suffix concatenates with ente
27:         Return stem
28:       end if
29:     else if word ends with suffix equals amente_suf then
30:       if Position of suffix in word is in R1 then
31:         Delete suffix in word
32:         stem_amente  $\leftarrow$  word without suffix
33:         if stem_amente preceded by iv_suf and the position of iv_suf is in R2 then

```

```

34:          Delete iv_suf in stem_amente
35:          stem_iv ← stem_amente without iv_suf
36:          if stem_iv further proceeded by at_suf and the position of at_suf is in
      R2 then
37:              Delete at_suf in stem_iv
38:              stem_at ← stem_iv without at_suf
39:              Return stem_at
40:          end if
41:          Return stem_iv
42:      else if stem_amente preceded by any suffix_oid in ozikad_suf then
43:          if Position of suffix_oid in stem_amente is in R2 then
44:              Delete suffix_oid in stem_amente
45:              stem_ozikad ← stem_amente without suffix_oid
46:              Return stem_ozikad
47:          end if
48:      end if
49:      Return stem_amente
50:  end if
51:  else if word ends with suffix equals mente_suf then
52:      if Position of suffix in word is in R2 then
53:          Delete suffix in word
54:          stem_mente ← word without suffix
55:          if stem_mente preceded by any suffix_ant in ante_suf then
56:              if Position of suffix_ant in stem_mente is in R2 then
57:                  Delete suffix_ant in stem_mente
58:                  stem_ante ← stem_mente without suffix_ant
59:                  Return stem_ante
60:              end if
61:          end if
62:          Return stem_mente
63:      end if
64:  else if word ends with suffix equals idade_suf then
65:      if Position of suffix in word is in R2 then
66:          Delete suffix in word
67:          stem_idade ← word without suffix
68:          if stem_idade preceded by any suffix_abl in abil_suf then
69:              if Position of suffix_abl in stem_idade is in R2 then
70:                  Delete suffix_abl in stem_idade
71:                  stem_abil ← stem_idade without suffix_abl
72:                  Return stem_abil

```

```

73:         end if
74:     end if
75:     Return stem_idade
76: end if
77: else if word ends with suffix equals iva_suf then
78:     if Position of suffix in word is in R2 then
79:         Delete suffix in word
80:         stem_iva ← word without suffix
81:         if stem_iva preceded by at_suf and the position of at_suf is in R2 then
82:             Delete at_suf in stem_iva
83:             stem_at ← stem_iva without at_suf
84:             Return stem_at
85:         end if
86:         Return stem_iva
87:     end if
88: else if word ends with suffix equals verb_suf then
89:     if Position of suffix in word is in RV then
90:         Delete suffix in word
91:         stem_verb ← word without suffix
92:         Return stem_verb
93:     end if
94: else if word ends with suffix equals residual_suf then
95:     if Position of suffix in word is in RV then
96:         Delete suffix in word
97:         stem_residual ← word without suffix
98:         Return stem_residual
99:     end if
100: else
101:     Return word
102: end if
103: end if
104: end for

```

▷ Step 3: Verb suffix removal

▷ Step 4: Residual suffix removal

A.2 Qualitative Assessment Criteria

The assessment criteria for the qualitative evaluation of the text corpus acquired through web crawling are presented in [Table A.3](#).

Table A.3: Assessment criteria used for the qualitative evaluation of the text corpus.

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, and 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.

A.3 Content Annotation Algorithm

The content annotation algorithm is outlined in Algorithm 1.

Algorithm 1 Content Annotation Algorithm.

Require: *start_text, end_text, documents, output_file*

```

1: for all document in documents do
2:   get title and url from document
3:   write title and url to output_file           ▷ The “annotated documents” file in Figure 4.3.
4:   get body_content from document
5:   annotation_t_counter  $\leftarrow$  0           ▷ To control the occurrence of < t > to a maximum of two.
6:   for all text_line in body_content do
7:     get text_line_lower by lowercasing text_line and removing extra spaces
8:     if text_line_lower starts with start_text and annotation_t_counter equals 0 then
9:       write annotation string < t >, a newline, text_line, and a newline to output_file
10:      Increment annotation_t_counter by 1
11:    else if text_line_lower ends with end_text and annotation_t_counter equals 1 then
12:      write text_line, a newline, annotation string < t >, and a newline to output_file
13:      Increment annotation_t_counter by 1
14:    else
15:      write text_line and a newline to output_file
16:    end if
17:  end for
18:  write an additional newline to output_file ▷ To separate each document by two newlines.
19: end for

```

A.4 Tetun Stopwords

The Tetun stopword list (*Labadain-Stopwords*), along with their English translations, is provided in [Table A.4](#).

Table A.4: Tetun stopwords list (*Labadain-Stopwords*) with English translations.

No.	Tetun	English	No.	Tetun	English	No.	Tetun	English	No.	Tetun	English
1	an	self	41	Hamutuk	with, together	81	laek	less	121	nunka	never
2	aleinde	besides	42	hanesan	such as	82	lai	for a while	122	o	you
3	ami	we	43	hela	remain	83	laiha	without	123	oin	next/sort/front
4	ami-nia	our	44	hikas	again	84	lalais	quickly	124	oin-oin	various
5	antes	before, previously	45	hira	how much	85	laran	inside	125	oinsá	how
6	atu	so that, to	46	hirak	those	86	leten	on, in	126	oioin	diverse
7	atubele	in order to	47	hirak-ne'e	these	87	liu	exceed, more than	127	oituan	few
8	ba	to, for	48	ho	with	88	liubá	ago	128	okos	below
9	baibain	usually, commonly	49	hodi	so that	89	liuhosi	through	129	ona	already
10	bainhira	when, while	50	hosi	from	90	liuhusi	through	130	ou	or
11	balu	some	51	hotu	too, also	91	liuliu	especially, particularly	131	para	to, in order to
12	barak	many	52	hotu-hotu	all	92	liután	further, more	132	portantu	so, therefore
13	bazeia	based (on), according (to)	53	husi	from	93	loloos	exactly, correctly	133	rasik	self, own
14	beibeik	often, always	54	i	and	94	loos	very, correct	134	resin	over, excess
15	bele	be able to, could, may	55	ida	a, an	95	lubuk	a lot of, many	135	ruma	some, any
16	besik	near, nearby, almost	56	ida-idak	each	96	mai	to, toward	136	sai	become, out
17	buat	thing	57	ida-ne'e	this	97	maibé	but	137	saida	what
18	dala	time(s)	58	ida-ne'ebé	which one	98	mais	however	138	se	if, whether
19	dalaruma	sometimes	59	iha	be, exist	99	maizumenus	more or less, approximately	139	sé	who
20	daudauk	currently	60	imi	you (plural)	100	mak	to be	140	sei	will, still
21	daudaun	currently	61	inklui	include	101	maka	to be	141	seidauk	not yet
22	de'it	only, just	62	ita	we	102	malu	each other	142	sein	without
23	depois	after that, then, later	63	ita-boot	you	103	mas	but	143	seluk	other
24	dezde	since, from	64	ita-nia	yours	104	maski	despite, although	144	sempre	always
25	didi'ak	carefully, thoroughly	65	ka	or	105	menus	less	145	sira	they
26	duke	than	66	kada	each, every	106	mezmu	despite, although	146	sira-ne'e	these
27	duni	indeed	67	karik	maybe	107	molok	before	147	sira-ne'ebé	those who
28	durante	during	68	katak	that	108	mós	also	148	sira-nia	their
29	eh	or	69	kedas	beforehand, immediately	109	nafatin	still, remain	149	sira-nian	theirs
30	enkuanu	while	70	komesa	from, begin	110	ne'e	this	150	só	only, unless
31	entaun	so, then	71	kona-ba	about	111	ne'ebá	that	151	tan	more
32	entre	between, among	72	kotuk	behind, last	112	ne'ebé	where	152	tanba	because
33	entretantu	meanwhile	73	kraik	below	113	nia	he, she	153	tantu	so
34	fali	again	74	kuandu	whenever, while	114	nian	of	154	tebes	very, so
35	filafali	again	75	kuaze	almost	115	ninia	his, her	155	tenke	must
36	foin	only just	76	la	not	116	ninian	his, hers	156	tiha	already
37	ha'u	I	77	la'ós	not	117	no	and	157	to'o	until
38	ha'u-nia	my	78	labele	unable, don't	118	nomós	also	158	tomak	entire
39	hafoin	after	79	ladún	not very, not so	119	nu'udar	as	159	tuur	according to
40	hahú	from, begin	80	lae	no	120	nune'e	like this, in this way	160	uitoan	few, a little

A.5 Search Topic Categories

The descriptions of the search topic categories are presented in [Table A.5](#).

Table A.5: Descriptions of search topic categories.

Topic	Description
(1) Education/Science	Queries related to educational topics, learning methods, school subjects, or courses.
(2) Health	Queries about physical or mental well-being, medical conditions, treatments, fitness, nutrition, or healthy lifestyle practices.
(3) Social/Cultural	Queries focus on societal norms, traditions, cultural practices, human behavior, or social issues.
(4) Research/Academia	Queries related to scholarly studies, research, scientific theories, or methodologies.
(5) Politic/Government	Queries about political systems and events or government structures, activities, and policies.
(6) Job vacancy	Queries about employment opportunities, including job openings, application processes, recruitment events, internships, or career advice.
(7) Economy/Finance	Queries focus on money management, investments, financial planning, economic trends, or market analysis.
(8) Agriculture/Environment	Queries concerning farming practices, environmental conservation, climate change, sustainability, or related topics.
(9) Law/Justice	Queries about legal principles, laws, regulations, court systems, individual rights, or justice procedures.
(10) Language/Translation	Queries related to language learning, translation, word definitions, grammar, or linguistic concepts.
(11) Computer science	Queries focused on computer science, software development, data analysis, information technology, or related topics.
(12) Personal	Queries of a personal nature, such as those about self-improvement, personal interests, hobbies, relationships, or private matters.
(13) Porn	Queries seeking explicit adult content, including pornography or other materials intended for adult audiences.
(14) Sports	Queries about sporting events, teams, athletes, rules, scores, or related topics.
(15) Holiday/Travel	Queries related to vacation planning, travel destinations, transportation, accommodation, or cultural experiences.
(16) Entertainment	focused on movies, music, television, celebrities, books, games, or other forms of leisure activities.
(17) Other	Queries that do not fall into any of the aforementioned categories or are difficult to interpret due to their brevity or related linguistic aspects.

References

- [1] Julien Abadji, Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. Towards a cleaner document-oriented multilingual crawled corpus. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odiijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 4344–4355. European Language Resources Association, 2022. URL <https://aclanthology.org/2022.lrec-1.463>.
- [2] Mirna Adriani, Jelita Asian, Bobby A. A. Nazief, Seyed M. M. Tahaghoghi, and Hugh E. Williams. Stemming indonesian: A confix-stripping approach. *ACM Transactions on Asian Language Information Processing*, 6(4):1–33, 2007. doi: 10.1145/1316457.1316459. URL <https://doi.org/10.1145/1316457.1316459>.
- [3] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, *VLDB’94, Proceedings of 20th International Conference on Very Large Data Bases, September 12-15, 1994, Santiago de Chile, Chile*, pages 487–499. Morgan Kaufmann, 1994. URL <http://www.vldb.org/conf/1994/P487.PDF>.
- [4] Rakesh Agrawal, Tomasz Imielinski, and Arun N. Swami. Mining association rules between sets of items in large databases. In Peter Buneman and Sushil Jajodia, editors, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington, DC, USA, May 26-28, 1993*, pages 207–216. ACM Press, 1993. doi: 10.1145/170035.170072. URL <https://doi.org/10.1145/170035.170072>.
- [5] Meta AI. Llama 3 model card. GitHub Repository, 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md. Accessed on April 22, 2024.
- [6] Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasajo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. One country, 700+ languages: NLP challenges for underrepresented languages and dialects in indonesia. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 7226–7249. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.ACL-LONG.500. URL <https://doi.org/10.18653/v1/2022.acl-long.500>.
- [7] Asadullah Al Galib, Md Humaion Kabir Mehedi, Ehsanur Rahman Rhythm, and Anna-jiat Alim Rasel. Large scale web crawling and distributed search engines: Techniques,

- challenges, current trends, and future prospects. In Nur Haryani Zakaria, Nur Suhaili Mansor, Husniza Husni, and Fathey Mohammed, editors, *Computing and Informatics*, pages 17–29, Singapore, 2024. Springer Nature Singapore. ISBN 978-981-99-9589-9. URL <http://doi.org/10.1007/978-981-99-9589-9>.
- [8] Abolfazl AleAhmad, Hadi Amiri, Ehsan Darrudi, Masoud Rahgozar, and Farhad Oroumchian. Hamshahri: A standard persian text collection. *Knowledge-Based Systems*, 22(5): 382–387, 2009. doi: 10.1016/J.KNOSYS.2009.05.002. URL <https://doi.org/10.1016/j.knosys.2009.05.002>.
- [9] Felermينو D. M. A. Ali, Andrew Caines, and Jaimito L. A. Malavi. Towards a parallel corpus of portuguese and the bantu language emakhuwa of mozambique. In Kathleen Siminyu, Julia Kreutzer, Hady Elsahar, Vukosi Marivate, Nishant Subramani, Jade Z. Abbott, and Bernardt Duvenhage, editors, *2nd AfricaNLP Workshop Proceedings, AfricaNLP@EACL 2021, Virtual Event, April 19, 2021*, 2021. URL <https://arxiv.org/abs/2104.05753>.
- [10] Felermينو D. M. A. Ali, Gabriel de Jesus, Henrique Lopes Cardoso, Sérgio Nunes, and Rui Sousa-Silva. Network-based approach for stopwords detection. In Pablo Gamallo, Daniela Claro, António Teixeira, Livy Real, Marcos Garcia, Hugo Gonçalo Oliveira, and Raquel Amaro, editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 2*, pages 55–63, Santiago de Compostela, Galicia/Spain, March 2024. Association for Computational Linguistics. URL <https://aclanthology.org/2024.propor-2.9>.
- [11] Gianni Amati and C. J. van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20(4):357–389, 2002. doi: 10.1145/582415.582416. URL <http://doi.acm.org/10.1145/582415.582416>.
- [12] Avishek Anand, Lawrence Cavedon, Matthias Hagen, Hideo Joho, Mark Sanderson, and Benno Stein. Dagstuhl seminar 19461 on conversational search: seminar goals and working group outcomes. *SIGIR Forum*, 54(1):3:1–3:11, 2020. doi: 10.1145/3451964.3451967. URL <https://doi.org/10.1145/3451964.3451967>.
- [13] S. Arie Ardiyanti, Dwi Hendratmo Widyantoro, Ayu Purwarianti, and Yayat Sudaryat. The rule-based sundanese stemmer. *ACM Transactions on Asian Language Information Processing*, 17(4):27:1–27:28, 2018. doi: 10.1145/3195634. URL <https://doi.org/10.1145/3195634>.
- [14] Jaime Arguello, Filip Radlinski, Hideo Joho, Damiano Spina, and Julia Kiseleva. Second international workshop on conversational approaches to information retrieval (cair’18): Workshop at SIGIR 2018. In Kevyn Collins-Thompson, Qiaozhu Mei, Brian D. Davison, Yiqun Liu, and Emine Yilmaz, editors, *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 1435–1437. ACM, 2018. doi: 10.1145/3209978.3210199. URL <https://doi.org/10.1145/3209978.3210199>.
- [15] Mark Aronoff. A decade of morphology and word formation. *Annual Review of Anthropology*, 12:355–375, 1983. URL <https://www.jstor.org/stable/2155652>.

- [16] Mikel Artetxe, Itziar Aldabe, Rodrigo Agerri, Olatz Perez-de-Viñaspre, and Aitor Soroa. Does corpus quality really matter for low-resource languages? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *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, 2022. doi: 10.18653/V1/2022.EMNLP-MAIN.499. URL <https://doi.org/10.18653/v1/2022.emnlp-main.499>.
- [17] Ricardo Baeza-Yates and Berthier A. Ribeiro-Neto. *Modern Information Retrieval - the concepts and technology behind search, Second edition*. Pearson Education Ltd., Harlow, England, 2011. ISBN 978-0-321-41691-9. URL <http://www.mir2ed.org/>.
- [18] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. Ms marco: A human generated machine reading comprehension dataset, 2018. URL <https://arxiv.org/abs/1611.09268>.
- [19] Protima Banerjee and Hyoil Han. Language modeling approaches to information retrieval. *J. Comput. Sci. Eng.*, 3(3):143–164, 2009. doi: 10.5626/jcse.2009.3.3.143. URL <https://doi.org/10.5626/jcse.2009.3.3.143>.
- [20] Marco Baroni and Silvia Bernardini. 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, 2004. URL <http://www.lrec-conf.org/proceedings/lrec2004/summaries/509.htm>.
- [21] A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences*, 101(11):3747–3752, 2004. doi: 10.1073/pnas.0400087101. URL <https://doi.org/10.1073/pnas.0400087101>.
- [22] Marc Barthelemy. Betweenness centrality in large complex networks. *The European Physical Journal B - Condensed Matter*, 38(2):163–168, mar 2004. doi: 10.1140/epjb/e2004-00111-4. URL <https://doi.org/10.1140%2Fepjb%2Fe2004-00111-4>.
- [23] Steven M. Beitzel, Eric C. Jensen, Abdur Chowdhury, David A. Grossman, and Ophir Frieder. Hourly analysis of a very large topically categorized web query log. In Mark Sanderson, Kalervo Järvelin, James Allan, and Peter Bruza, editors, *SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, UK, July 25-29, 2004*, pages 321–328. ACM, 2004. doi: 10.1145/1008992.1009048. URL <https://doi.org/10.1145/1008992.1009048>.
- [24] Adam L. Berger, Rich Caruana, David Cohn, Dayne Freitag, and Vibhu O. Mittal. Bridging the lexical chasm: statistical approaches to answer-finding. In Emmanuel J. Yannakoudakis, Nicholas J. Belkin, Peter Ingwersen, and Mun-Kew Leong, editors, *SIGIR 2000: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 24-28, 2000, Athens, Greece*, pages 192–199. ACM, 2000. doi: 10.1145/345508.345576. URL <https://doi.org/10.1145/345508.345576>.

- [25] Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. Automatic speech recognition for under-resourced languages: A survey. *Speech Commun.*, 56:85–100, 2014. doi: 10.1016/J.SPECOM.2013.07.008. URL <https://doi.org/10.1016/j.speechom.2013.07.008>.
- [26] Verena Blaschke, Hinrich Schütze, and Barbara Plank. A survey of corpora for germanic low-resource languages and dialects. In Tanel Alumäe and Mark Fishel, editors, *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, 2023. URL <https://aclanthology.org/2023.nodalida-1.41>.
- [27] Luiz Henrique Bonifacio, Israel Campiotti, Roberto A. Lotufo, and Rodrigo Nogueira. mmarco: A multilingual version of MS MARCO passage ranking dataset. *CoRR*, abs/2108.13897, 2021. URL <https://arxiv.org/abs/2108.13897>.
- [28] Ulrik Brandes. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2):163–177, 2001. doi: 10.1080/0022250X.2001.9990249. URL <https://doi.org/10.1080/0022250X.2001.9990249>.
- [29] Martin Braschler and Bärbel Ripplinger. How effective is stemming and compounding for german text retrieval? *Information Retrieval*, 7(3-4):291–316, 2004. doi: 10.1023/B:INRT.0000011208.60754.A1. URL <https://doi.org/10.1023/B:INRT.0000011208.60754.a1>.
- [30] The Editors of Encyclopaedia Britannica. Assamese language. Encyclopedia Britannica, 2024. URL <https://www.britannica.com/topic/Assamese-language>. Accessed on February 19, 2024.
- [31] Ralf D. Brown. Selecting and weighting n-grams to identify 1100 languages. In Ivan Habernal and Václav Matousek, editors, *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, 2013. doi: 10.1007/978-3-642-40585-3_60. URL https://doi.org/10.1007/978-3-642-40585-3_60.
- [32] Jay Budzik, Kristian J. Hammond, and Lawrence Birnbaum. Information access in context. *Knowl. Based Syst.*, 14(1-2):37–53, 2001. doi: 10.1016/S0950-7051(00)00105-2. URL [https://doi.org/10.1016/S0950-7051\(00\)00105-2](https://doi.org/10.1016/S0950-7051(00)00105-2).
- [33] Mirelle Bueno, Eduardo Seiti de Oliveira, Rodrigo Frassetto Nogueira, Roberto A. Lotufo, and Jayr Alencar Pereira. Quati: A brazilian portuguese information retrieval dataset from native speakers. *CoRR*, abs/2404.06976, 2024. doi: 10.48550/ARXIV.2404.06976. URL <https://doi.org/10.48550/arXiv.2404.06976>.
- [34] Vannevar Bush. As we may think. *Atlantic Monthly*, 176(1):101–108, 1945. URL <http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm>.
- [35] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. Learning to rank: from pairwise approach to listwise approach. In Zoubin Ghahramani, editor, *Machine Learning, Proceedings of the Twenty-Fourth International Conference (ICML 2007), Corvallis, Oregon, USA, June 20-24, 2007*, volume 227 of *ACM International Conference Proceeding Series*, pages 129–136. ACM, 2007. doi: 10.1145/1273496.1273513. URL <https://doi.org/10.1145/1273496.1273513>.

- [36] Catherine Chavula. *Using language similarities in retrieval for resource scarce languages: a study of several southern Bantu language*. PhD thesis, University of Cape Town, Cape Town, South Africa, 2021. URL <http://hdl.handle.net/11427/33614>.
- [37] Catherine Chavula and Hussein Suleman. Ranking by language similarity for resource scarce southern bantu languages. In Faegheh Hasibi, Yi Fang, and Akiko Aizawa, editors, *ICTIR '21: The 2021 ACM SIGIR International Conference on the Theory of Information Retrieval, Virtual Event, Canada, July 11, 2021*, pages 137–147. ACM, 2021. doi: 10.1145/3471158.3472251. URL <https://doi.org/10.1145/3471158.3472251>.
- [38] Heng Chen, Xinying Chen, and Haitao Liu. How does language change as a lexical network? an investigation based on written chinese word co-occurrence networks. *PLOS ONE*, 13(2):1–22, 02 2018. doi: 10.1371/journal.pone.0192545. URL <https://doi.org/10.1371/journal.pone.0192545>.
- [39] Tao Chen, Mingyang Zhang, Jing Lu, Michael Bendersky, and Marc Najork. Out-of-domain semantics to the rescue! zero-shot hybrid retrieval models. In Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty, editors, *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part I*, volume 13185 of *Lecture Notes in Computer Science*, pages 95–110. Springer, 2022. doi: 10.1007/978-3-030-99736-6_7. URL https://doi.org/10.1007/978-3-030-99736-6_7.
- [40] Christopher Cieri, Mike Maxwell, Stephanie M. Strassel, and Jennifer Tracey. Selection criteria for low resource language programs. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA), 2016. URL <http://www.lrec-conf.org/proceedings/lrec2016/summaries/1254.html>.
- [41] Cyril Cleverdon. The cranfield tests on index language devices. In *Aslib proceedings*, volume 19, pages 173–194. MCB UP Ltd, 1967.
- [42] Paul D. Clough and Mark Sanderson. Evaluating the performance of information retrieval systems using test collections. *Information Resesearch*, 18(2), 2013. URL <http://www.informationr.net/ir/18-2/paper582.html>.
- [43] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46, 1960. doi: 10.1177/001316446002000104. URL <https://doi.org/10.1177/001316446002000104>.
- [44] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.ACL-MAIN.747. URL <https://doi.org/10.18653/v1/2020.acl-main.747>.

- [45] Conselho de Imprensa Timor-Leste. Relatório resultado levantamento nacional: “instituição mídia profissional no rekursu jornalista iha timor-leste”. Technical Report CI Nú. 01, Conselho de Imprensa Timor-Leste, May 2017. URL <https://conselhoimprensa.tl/baze-de-dadus/peskiza/download/13-peskiza/65-ci>.
- [46] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava. Data preparation for mining world wide web browsing patterns. *Knowl. Inf. Syst.*, 1(1):5–32, 1999. doi: 10.1007/BF03325089. URL <https://doi.org/10.1007/BF03325089>.
- [47] Alissa Cooper. A survey of query log privacy-enhancing techniques from a policy perspective. *ACM Trans. Web*, 2(4):19:1–19:27, 2008. doi: 10.1145/1409220.1409222. URL <https://doi.org/10.1145/1409220.1409222>.
- [48] Gordon V. Cormack, Charles L. A. Clarke, and Stefan Büttcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In James Allan, Javed A. Aslam, Mark Sanderson, ChengXiang Zhai, and Justin Zobel, editors, *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009*, pages 758–759. ACM, 2009. doi: 10.1145/1571941.1572114. URL <https://doi.org/10.1145/1571941.1572114>.
- [49] Adérito José Guterres Correia, Geoffrey Stephen Hull, George William Saunders, and Domingos dos Santos Rosa da Costa Tilman, Mário Adriano Soares. *Disionáriu Nasionál ba Tetun Ofisiál*. Instituto Nacional de Linguística, Universidade Nacional Timor Lorosa’e, Avenida Cidade de Lisboa, Dili, Timor-Leste, 2005. ISBN 1-74138-160-6.
- [50] Nick Craswell. Mean reciprocal rank. In Ling Liu and M. Tamer Özsu, editors, *Encyclopedia of Database Systems*, page 1703. Springer US, 2009. doi: 10.1007/978-0-387-39940-9_488. URL https://doi.org/10.1007/978-0-387-39940-9_488.
- [51] Nick Craswell. Mean reciprocal rank. In Ling Liu and M. Tamer Özsu, editors, *Encyclopedia of Database Systems, Second Edition*. Springer, 2018. doi: 10.1007/978-1-4614-8265-9_488. URL https://doi.org/10.1007/978-1-4614-8265-9_488.
- [52] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. Overview of the trec 2019 deep learning track. In *Proceedings of The Twenty-Eighth Text REtrieval Conference (TREC 2019)*, 2019. URL <https://trec.nist.gov/pubs/trec28/papers/OVERVIEW.DL.pdf>.
- [53] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Jimmy Lin. Overview of the TREC 2021 deep learning track. In Ian Soboroff and Angela Ellis, editors, *Proceedings of the Thirtieth Text REtrieval Conference, TREC 2021, online, November 15-19, 2021*, volume 500-335 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2021. URL <https://trec.nist.gov/pubs/trec30/papers/Overview-DL.pdf>.
- [54] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Hossein A. Rahmani, Daniel Campos, Jimmy Lin, Ellen M. Voorhees, and Ian Soboroff. Overview of the TREC 2023 deep learning track. In Ian Soboroff and Angela Ellis, editors, *The Thirty-Second Text REtrieval Conference Proceedings (TREC 2023)*, Gaithersburg, MD, USA, November 14-17, 2023, volume 500-xxx of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2023. URL <https://trec.nist.gov/pubs/trec32/papers/Overview-deep.pdf>.

- [55] W. Bruce Croft, Donald Metzler, and Trevor Strohman. *Search Engines - Information Retrieval in Practice*. Pearson Education, 2009. ISBN 978-0-13-136489-9. URL <http://www.search-engines-book.com/>.
- [56] Jornal da República de Timor-Leste. Deliberação N.º 20/2020, de 30 de Outubro: Aprovação do pedido de Registo da Sociedade por quotas “BTN (Borulili Timor News), Unipessoal, Lda” como Órgão de comunicação social, 2020. Série I, N.º 45, https://www.mj.gov.tl/jornal/public/docs/2020/serie_1/SERIE_I_NO_45.pdf.
- [57] Instituto Nacional de Estatística Timor-Leste (INETL). Census 2015 priority table population by language, 2015. URL <https://inetl-ip.gov.tl/2023/03/09/census-2015-priority-table-population-by-language/>. Accessed on November 19, 2024.
- [58] Instituto Nacional de Estatística Timor-Leste (INETL). Timor-Leste population and housing census, 2022. URL <https://inetl-ip.gov.tl/2023/10/04/2022-census-wall-chart/>. Accessed on November 8, 2024.
- [59] Gabriel de Jesus. Timor News: An online news portal exclusively publishing content in Tetun. Online news portal, 2019. URL <https://www.timornews.tl>.
- [60] Gabriel de Jesus. Text information retrieval in Tetun. In Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo, editors, *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, Switzerland, 2023. Springer. doi: 10.1007/978-3-031-28241-6_48. URL https://doi.org/10.1007/978-3-031-28241-6_48.
- [61] Gabriel de Jesus. Text information retrieval in Tetun: A preliminary study. *CoRR*, abs/2406.07331, 2024. doi: 10.48550/ARXIV.2406.07331. URL <https://doi.org/10.48550/arXiv.2406.07331>. This work was published on the 10th edition of the PhD Symposium on FDIA, July 20, 2022, Lisbon, Portugal.
- [62] Gabriel de Jesus and Sérgio Nunes. Labadain-30k+: A monolingual Tetun document-level audited dataset. In Maite Melero, Sakriani Sakti, and Claudia Soria, editors, *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 177–188, Torino, Italia, May 2024. ELRA and ICCL. URL <https://aclanthology.org/2024.sigul-1.22>.
- [63] Gabriel de Jesus and Sérgio Nunes. Data collection pipeline for low-resource languages: A case study on constructing a Tetun text corpus [crawler tool]. Zenodo, 2024. URL <https://doi.org/10.5281/zenodo.10911381>.
- [64] Gabriel de Jesus and Sérgio Nunes. Labadain-30k+: A monolingual Tetun document-level audited dataset [data set]. INESC TEC, 2024. URL <https://doi.org/10.25747/YDWR-N696>.
- [65] Gabriel de Jesus and Sérgio Nunes. Exploring large language models for relevance judgments in Tetun. In C. Siro, M. Aliannejadi, H.A. Rahmani, N. Craswell, C.L.A. Clarke, G. Faggioli, B. Mitra, P. Thomas, and E. Yilmaz, editors, *Proceedings of The First Workshop on Large Language Models for Evaluation in Information Retrieval (LLM4Eval*

- 2024), co-located with the 10th International Conference on Online Publishing (SIGIR 2024), volume 3752, pages 19–30, Washington D.C., USA, July 2024. URL <https://ceur-ws.org/Vol-3752/>.
- [66] Gabriel de Jesus and Sérgio Nunes. Tetun language identification model: A python package for Tetun LID. PyPI Package, 2024. URL <https://pypi.org/project/tetun-lid/>.
- [67] Gabriel de Jesus and Sérgio Nunes. Tetun tokenizer: A python package for tokenizing Tetun text. PyPI Package, 2024. URL <https://pypi.org/project/tetun-tokenizer/>.
- [68] Gabriel de Jesus and Sérgio Nunes. Labadain Chat: An LLM-based conversational system designed for Tetun speakers. Online web chat, 2024. URL <https://www.labadain.com>.
- [69] Gabriel de Jesus and Sérgio Nunes. Labadain Search: A monolingual search engine developed for Tetun speakers. Online web search engine, 2024. URL <https://www.labadain.tl>.
- [70] Gabriel de Jesus and Sérgio Nunes. Labadain-Avaliadór: A Test Collection for Tetun Ad-hoc Text Retrieval Task [dataset]. INESC TEC, 2025. URL <https://doi.org/10.25747/2k6s-e518>.
- [71] Gabriel de Jesus and Sérgio Nunes. Labadainlog-17k+: Search logs from Tetun-speaking users across chat, web, and news platforms [dataset]. INESC TEC, 2025. URL <https://doi.org/10.25747/x2dk-5y06>.
- [72] Gabriel de Jesus and Sérgio Nunes. Labadain-Stopwords: A curated list of 160 Tetun stopwords [dataset]. INESC TEC, 2025. URL <https://doi.org/10.25747/pg2v-kx70>.
- [73] Gabriel de Jesus and Sérgio Nunes. Establishing a foundation for Tetun ad-hoc text retrieval: Stemming, indexing, retrieval, and ranking, 2025. URL <https://arxiv.org/abs/2412.11758>. The paper has been submitted to the ACM TALLIP journal and is currently under peer review.
- [74] Gabriel de Jesus and Sérgio Sobral Nunes. Data collection pipeline for low-resource languages: A case study on constructing a Tetun text corpus. In Nicoletta Calzolari, Min-Yen Kan, Véronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 4368–4380. ELRA and ICCL, 2024. URL <https://aclanthology.org/2024.lrec-main.390>.
- [75] Gabriel de Jesus and Sérgio Nunes. Labadain-Stemmer: A stemming algorithm designed for the Tetun language. GitHub Repository, 2025. URL <https://github.com/gabriel-de-jesus/labadain-stemmer>.
- [76] Gabriel de Jesus, Siddharth AK Singh, Sérgio Nunes, and Andrew Yates. Labadain-ZSRunS: A sparse and zero-shot dense retrieval runs with LLM-generated summaries for Tetun ad-hoc text retrieval [data set]. INESC TEC, 2025. URL <https://doi.org/10.25747/rfzx-m945>.

- [77] Gabriel de Jesus, Siddharth A.K. Singh, Sérgio Nunes, and Andrew Yates. Zero-Shot and Hybrid Strategies for Tetun Ad-Hoc Text Retrieval. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, ICTIR '25, page 264–274, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400718618. doi: 10.1145/3731120.3744593. URL <https://doi.org/10.1145/3731120.3744593>.
- [78] Instituto Nacional de Linguística. *Matadalan ortográfiku ba Tetun ofisiál*. Instituto Nacional de Linguística), Universidade Nacional Timor Lorosa'e, Avenida Cidade de Lisboa, Dili, Timor-Leste, 2006. ISBN 9781864087123.
- [79] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/V1/N19-1423. URL <https://doi.org/10.18653/v1/n19-1423>.
- [80] Ljiljana Dolamic and Jacques Savoy. Indexing and stemming approaches for the czech language. *Information Processing & Management*, 45(6):714–720, 2009. doi: 10.1016/J.IPM.2009.06.001. URL <https://doi.org/10.1016/j.ipm.2009.06.001>.
- [81] Ljiljana Dolamic and Jacques Savoy. Comparative study of indexing and search strategies for the hindi, marathi, and bengali languages. *ACM Transactions on Asian Language Information Processing*, 9(3):11:1–11:24, 2010. doi: 10.1145/1838745.1838748. URL <https://doi.org/10.1145/1838745.1838748>.
- [82] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. *CoRR*, abs/2401.08281, 2024. doi: 10.48550/ARXIV.2401.08281. URL <https://doi.org/10.48550/arXiv.2401.08281>.
- [83] Olha Dovbnia, Witold Sosnowski, and Anna Wróblewska. Automatic language identification for celtic texts. In Mohammad Tanveer, Sonali Agarwal, Seiichi Ozawa, Asif Ekbal, and Adam Jatowt, editors, *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, 2022. doi: 10.1007/978-981-99-1645-0_22. URL https://doi.org/10.1007/978-981-99-1645-0_22.
- [84] Carsten Eickhoff, Jaime Teevan, Ryen White, and Susan T. Dumais. Lessons from the journey: a query log analysis of within-session learning. In Ben Carterette, Fernando Diaz, Carlos Castillo, and Donald Metzler, editors, *Seventh ACM International Conference on Web Search and Data Mining, WSDM 2014, New York, NY, USA, February 24-28, 2014*, pages 223–232. ACM, 2014. doi: 10.1145/2556195.2556217. URL <https://doi.org/10.1145/2556195.2556217>.
- [85] Alexandra Espichán-Linares and Arturo Oncevay-Marcos. Language identification with scarce data: A case study from peru. In Juan Antonio Lossio-Ventura and Hugo Alatrasta-Salas, editors, *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, 2017. doi: 10.1007/978-3-319-90596-9_7. URL https://doi.org/10.1007/978-3-319-90596-9_7.
- [86] Alexandra Espichán-Linares and Arturo Oncevay-Marcos. Language identification with scarce data: A case study from peru. In Juan Antonio Lossio-Ventura and Hugo Alatrística-Salas, editors, *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, 2017. doi: 10.1007/978-3-319-90596-9_7. URL https://doi.org/10.1007/978-3-319-90596-9_7.
- [87] Kobra Etminani, Mohammad R. Akbarzadeh-Totonchi, and Noorali Raeji Yanehsari. Web usage mining: users’ navigational patterns extraction from web logs using ant-based clustering method. In João Paulo Carvalho, Didier Dubois, Uzay Kaymak, and João Miguel da Costa Sousa, editors, *Proceedings of the Joint 2009 International Fuzzy Systems Association World Congress and 2009 European Society of Fuzzy Logic and Technology Conference, Lisbon, Portugal, July 20-24, 2009*, pages 396–401, 2009. URL http://www.eusflat.org/proceedings/IFSA-EUSFLAT_2009/pdf/tema_0396.pdf.
- [88] Leonhard Euler. Solution of a problem in the geometry of position, 1736. URL https://www.cantab.net/users/michael.behrend/repubs/maze_maths/pages/euler_en.html. Translated version available online. Accessed on November 8, 2024.
- [89] Guglielmo Faggioli, Laura Dietz, Charles L. A. Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. Perspectives on large language models for relevance judgment. In Masaharu Yoshioka, Julia Kiseleva, and Mohammad Aliannejadi, editors, *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2023, Taipei, Taiwan, 23 July 2023*, pages 39–50. ACM, 2023. doi: 10.1145/3578337.3605136. URL <https://doi.org/10.1145/3578337.3605136>.
- [90] Yixing Fan, Xiaohui Xie, Yinqiong Cai, Jia Chen, Xinyu Ma, Xiangsheng Li, Ruqing Zhang, and Jiafeng Guo. Pre-training methods in information retrieval. *Found. Trends Inf. Retr.*, 16(3):178–317, 2022. doi: 10.1561/1500000100. URL <https://doi.org/10.1561/1500000100>.
- [91] Stefano Ferilli. Automatic multilingual stopwords identification from very small corpora. *Electronics*, 10(17), 2021. ISSN 2079-9292. doi: 10.3390/electronics10172169. URL <https://www.mdpi.com/2079-9292/10/17/2169>.
- [92] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382, 1971. URL <https://api.semanticscholar.org/CorpusID:143544759>. Measures of inter-rater reliability for two or more rates (annotators).
- [93] Felipe N. Flores and Viviane Pereira Moreira. Assessing the impact of stemming accuracy on information retrieval - A multilingual perspective. *Information Processing & Management*, 52(5):840–854, 2016. doi: 10.1016/J.IPM.2016.03.004. URL <https://doi.org/10.1016/j.ipm.2016.03.004>.
- [94] Christopher J. Fox. A stop list for general text. *SIGIR Forum*, 24(1-2):19–35, 1990. doi: 10.1145/378881.378888. URL <https://doi.org/10.1145/378881.378888>.

- [95] William B. Frakes and Christopher J. Fox. Strength and similarity of affix removal stemming algorithms. *SIGIR Forum*, 37(1):26–30, apr 2003. ISSN 0163-5840. doi: 10.1145/945546.945548. URL <https://doi.org/10.1145/945546.945548>.
- [96] Eibe Frank, Mark A. Hall, and Ian H. Witten. *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*. Morgan Kaufmann, fourth edition, 2016. URL https://ml.cms.waikato.ac.nz/weka/Witten_et_al_2016_appendix.pdf.
- [97] Luyu Gao and Jamie Callan. Unsupervised corpus aware language model pre-training for dense passage retrieval. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2022, Dublin, Ireland, May 22-27, 2022, pages 2843–2853. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.ACL-LONG.203. URL <https://doi.org/10.18653/v1/2022.acl-long.203>.
- [98] Luyu Gao, Zhuyun Dai, Tongfei Chen, Zhen Fan, Benjamin Van Durme, and Jamie Callan. Complement lexical retrieval model with semantic residual embeddings. In Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani, editors, *Advances in Information Retrieval - 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 - April 1, 2021, Proceedings, Part I*, volume 12656 of *Lecture Notes in Computer Science*, pages 146–160. Springer, 2021. doi: 10.1007/978-3-030-72113-8_10. URL https://doi.org/10.1007/978-3-030-72113-8_10.
- [99] Yuyang Gao, Wei Liang, Yuming Shi, and Qiuling Huang. Comparison of directed and weighted co-occurrence networks of six languages. *Physica A: Statistical Mechanics and its Applications*, 393:579–589, 2014. ISSN 0378-4371. doi: <https://doi.org/10.1016/j.physa.2013.08.075>. URL <https://www.sciencedirect.com/science/article/pii/S037843711300825X>.
- [100] Kripabandhu Ghosh and Arnab Bhattacharya. Stopword removal: Why bother? A case study on verbose queries. In Partha Pratim Chakraborty, Manish Gupta, Lipika Dey, and Shourya Roy, editors, *Proceedings of the 10th Annual ACM India Compute Conference, Compute 2017, Bhopal, India, November 16-18, 2017*, pages 99–102. ACM, 2017. doi: 10.1145/3140107.3140125. URL <https://doi.org/10.1145/3140107.3140125>.
- [101] Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 759–765. European Language Resources Association (ELRA), 2012. URL <http://www.lrec-conf.org/proceedings/lrec2012/summaries/327.html>.
- [102] Edouard Grave, Piotr Bojanowski, Prakhhar Gupta, Armand Joulin, and Tomáš Mikolov. Learning word vectors for 157 languages. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Kôiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources

- Association (ELRA), 2018. URL <http://www.lrec-conf.org/proceedings/lrec2018/summaries/627.html>.
- [103] Zuzana Greksáková. *Tetun in Timor-Leste: The role of language contact in its development*. PhD thesis, Universidade de Coimbra, Portugal, 2018. URL <http://hdl.handle.net/10316/80665>.
- [104] Geoffrey C. Gunn. *Timor Loro Sae: 500 years*. Livros do Oriente, 1999. ISBN 972-9418-69-1. URL <https://nla.gov.au/nla.cat-vn3789875>.
- [105] Jiafeng Guo, Yinqiong Cai, Yixing Fan, Fei Sun, Ruqing Zhang, and Xueqi Cheng. Semantic models for the first-stage retrieval: A comprehensive review. *ACM Trans. Inf. Syst.*, 40(4):66:1–66:42, 2022. doi: 10.1145/3486250. URL <https://doi.org/10.1145/3486250>.
- [106] John Hajek and Catharina Williams van Klinken. Language contact and gender in Tetun Dili: What happens when austronesian meets romance? *Oceanic Linguistics*, 58:59–91, 06 2019. doi: 10.1353/ol.2019.0003.
- [107] Donna K. Harman, editor. *Overview of the First Text REtrieval Conference (TREC-1)*, volume 500-207 of *NIST Special Publication*, 1992. National Institute of Standards and Technology (NIST). URL http://trec.nist.gov/pubs/trec1/t1_proceedings.html.
- [108] David Hawking. Overview of the TREC-9 web track. In Ellen M. Voorhees and Donna K. Harman, editors, *Proceedings of The Ninth Text REtrieval Conference, TREC 2000, Gaithersburg, Maryland, USA, November 13-16, 2000*, volume 500-249 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2000. URL <http://trec.nist.gov/pubs/trec9/papers/web9.pdf>.
- [109] Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. A survey on recent approaches for natural language processing in low-resource scenarios. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *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, 2021. doi: 10.18653/V1/2021.NAACL-MAIN.201. URL <https://doi.org/10.18653/v1/2021.naacl-main.201>.
- [110] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL <https://openreview.net/forum?id=d7KBjmI3GmQ>.
- [111] Djoerd Hiemstra. *Using Language Models for Information Retrieval*. PhD thesis, University of Twente, Enschede, Netherlands, 2001. URL <https://ris.utwente.nl/ws/files/6042641/>.
- [112] Djoerd Hiemstra and Wessel Kraaij. Twenty-one at trec-7: ad-hoc and cross-language track. In E.M. Voorhees and D.K. Harman, editors, *Proceedings of the seventh Text Retrieval Conference (TREC)*, NIST Special Publications, pages 227–238, United States, 1999. National

- Institute of Standards and Technology. URL https://trec.nist.gov/pubs/trec7/t7_proceedings.html.
- [113] Jun Hirai, Sriram Raghavan, Hector Garcia-Molina, and Andreas Paepcke. Webbase: a repository of web pages. *Computer Networks*, 33(1-6):277–293, 2000. doi: 10.1016/S1389-1286(00)00063-3. URL [https://doi.org/10.1016/S1389-1286\(00\)00063-3](https://doi.org/10.1016/S1389-1286(00)00063-3).
- [114] Armin Hoenen, Cemre Koc, and Marc Daniel Rahn. A manual for web corpus crawling of low resource languages. *Umanistica Digitale*, 4(8):342–344, Jan. 2020. doi: 10.6092/issn.2532-8816/9931. URL <https://umanisticadigitale.unibo.it/article/view/9931>.
- [115] Vera Hollink, Jaap Kamps, Christof Monz, and Maarten de Rijke. Monolingual document retrieval for european languages. *Information Retrieval*, 7(1-2):33–52, 2004. doi: 10.1023/B:INRT.0000009439.19151.4C. URL <https://doi.org/10.1023/B:INRT.0000009439.19151.4c>.
- [116] Geoffrey Stephen Hull and Adérito José Guterres Correia. *Kursu Gramátika Tetun ba Profesór, Tradutór, Jornalista, no Estudante-Universidade Sira*. Instituto Nacional de Linguística (INL), 2005. ISBN 1-7413-8137-1.
- [117] International Organization for Standardization. ISO 8601: Date and time format, 2004. URL <https://www.iso.org/iso-8601-date-and-time-format.html>. Accessed on January 22, 2024.
- [118] International Telecommunication Union (ITU). Individuals using the internet (% of population) - Timor-Leste, 2024. URL <https://data.worldbank.org/indicator/IT.NET.USER.ZS?locations=TL>. World Telecommunication/ICT Indicators Database, accessed on January 22, 2024.
- [119] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *Trans. Mach. Learn. Res.*, 2022, 2022. URL <https://openreview.net/forum?id=jKN1pXi7b0>.
- [120] Bernard J. Jansen, Danielle L. Booth, and Amanda Spink. Determining the informational, navigational, and transactional intent of web queries. *Inf. Process. Manag.*, 44(3):1251–1266, 2008. doi: 10.1016/J.IPM.2007.07.015. URL <https://doi.org/10.1016/j.ipm.2007.07.015>.
- [121] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002. doi: 10.1145/582415.582418. URL <http://doi.acm.org/10.1145/582415.582418>.
- [122] Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research*, 65:675–782, 2019. doi: 10.1613/JAIR.1.11675. URL <https://doi.org/10.1613/jair.1.11675>.
- [123] Nour Jedidi, Yung-Sung Chuang, Leslie Shing, and James R. Glass. Zero-shot dense retrieval with embeddings from relevance feedback. *CoRR*, abs/2410.21242, 2024. doi:

- 10.48550/ARXIV.2410.21242. URL <https://doi.org/10.48550/arXiv.2410.21242>.
- [124] Frederick Jelinek and Robert Leroy Mercer. Interpolated estimation of Markov source parameters from sparse data. In *Proceedings of the Workshop on Pattern Recognition in Practice*, pages 381–397, Amsterdam, 1980. North Holland.
- [125] Gabriel de Jesus and Sérgio Nunes. Insights into LLM-Based Conversational Search: A Study of Tetun-Speaking Users’ Search Behavior. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, ICTIR ’25, page 297–306, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400718618. doi: 10.1145/3731120.3744596. URL <https://doi.org/10.1145/3731120.3744596>.
- [126] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6769–6781. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.EMNLP-MAIN.550. URL <https://doi.org/10.18653/v1/2020.emnlp-main.550>.
- [127] Jaana Kekäläinen. Binary and graded relevance in IR evaluations—comparison of the effects on ranking of IR systems. *Information Processing & Management*, 41(5):1019–1033, 2005. doi: 10.1016/J.IPM.2005.01.004. URL <https://doi.org/10.1016/j.ipm.2005.01.004>.
- [128] Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over BERT. In Jimmy X. Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu, editors, *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 39–48. ACM, 2020. doi: 10.1145/3397271.3401075. URL <https://doi.org/10.1145/3397271.3401075>.
- [129] Ekaterina Khramtsova, Shengyao Zhuang, Mahsa Baktashmotlagh, Xi Wang, and Guido Zuccon. Selecting which dense retriever to use for zero-shot search. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, SIGIR-AP ’23*, page 223–233, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400704086. doi: 10.1145/3624918.3625330. URL <https://doi.org/10.1145/3624918.3625330>.
- [130] Hyunwoo Kim, Yoonseo Choi, Taehyun Yang, Honggu Lee, Chaneon Park, Yongju Lee, Jin Young Kim, and Juho Kim. Using llms to investigate correlations of conversational follow-up queries with user satisfaction. In C. Siro, M. Aliannejadi, H.A. Rahmani, N. Craswell, C.L.A. Clarke, G. Faggioli, B. Mitra, P. Thomas, and E. Yilmaz, editors, *Proceedings of The First Workshop on Large Language Models for Evaluation in Information Retrieval (LLM4Eval 2024), co-located with the 10th International Conference on Online Publishing (SIGIR 2024)*, volume 3752, pages 66–91, Washington D.C., USA, July 2024. URL <https://ceur-ws.org/Vol-3752/>.

- [131] Ben King and Steven P. Abney. Labeling the languages of words in mixed-language documents using weakly supervised methods. In Lucy Vanderwende, Hal Daumé III, and Kartrin Kirchhoff, editors, *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 1110–1119. The Association for Computational Linguistics, 2013. URL <https://aclanthology.org/N13-1131/>.
- [132] Philipp Koehn, Francisco Guzmán, Vishrav Chaudhary, and Juan Miguel Pino. Findings of the WMT 2019 shared task on parallel corpus filtering for low-resource conditions. In Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno-Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névél, Mariana L. Neves, Matt Post, Marco Turchi, and Karin Verspoor, editors, *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, 2019. doi: 10.18653/V1/W19-5404. URL <https://doi.org/10.18653/v1/w19-5404>.
- [133] Erik Körner, Felix Helfer, Christopher Schröder, Thomas Eckart, and Dirk Goldhahn. 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), 2022. URL <https://aclanthology.org/2022.dclrl-1.5.pdf>.
- [134] 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, Irero Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhlov, 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. Quality at a glance: An audit of web-crawled multilingual datasets. *Trans. Assoc. Comput. Linguistics*, 10:50–72, 2022. doi: 10.1162/TACL_A_00447. URL https://doi.org/10.1162/tacl_a_00447.
- [135] Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 66–75. Association for Computational Linguistics, 2018. doi: 10.18653/V1/P18-1007. URL <https://aclanthology.org/P18-1007/>.
- [136] Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. MADLAD-400: A multilingual and document-level large audited dataset. In Alice Oh, Tristan Naumann, Amir Globerson,

- Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/d49042a5d49818711c401d34172f9900-Abstract-Datasets_and_Benchmarks.html.
- [137] Siva Subrahmanyam Varma Kusampudi, Anudeep Chaluvadi, and Radhika Mamidi. Corpus creation and language identification in low-resource code-mixed telugu-english text. In Galia Angelova, Maria Kuniolkovskaya, Ruslan Mitkov, and Ivelina Nikolova-Koleva, editors, *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., 2021. URL <https://aclanthology.org/2021.ranlp-1.85>.
- [138] Saar Kuzi, Mingyang Zhang, Cheng Li, Michael Bendersky, and Marc Najork. Leveraging semantic and lexical matching to improve the recall of document retrieval systems: A hybrid approach. *CoRR*, abs/2010.01195, 2020. URL <https://arxiv.org/abs/2010.01195>.
- [139] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33 1:159–74, 1977. URL <https://api.semanticscholar.org/CorpusID:11077516>. The reference contains interpretation of k-value of inter-annotators. The interpretation is only for two annotators and two class. It is used in interpreting Fleiss’ Kappa.
- [140] Dawn J. Lawrie, Eugene Yang, Douglas W. Oard, and James Mayfield. Neural approaches to multilingual information retrieval. In Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo, editors, *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part I*, volume 13980 of *Lecture Notes in Computer Science*, pages 521–536. Springer, 2023. doi: 10.1007/978-3-031-28244-7_33. URL https://doi.org/10.1007/978-3-031-28244-7_33.
- [141] Dohyeon Lee, Seung-won Hwang, Kyungjae Lee, Seungtaek Choi, and Sunghyun Park. On complementarity objectives for hybrid retrieval. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13357–13368. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.746. URL <https://doi.org/10.18653/v1/2023.acl-long.746>.
- [142] Wei Liang, Yuming Shi, Chi K. Tse, Jing Liu, Yanli Wang, and Xunqiang Cui. Comparison of co-occurrence networks of the chinese and english languages. *Physica A: Statistical Mechanics and its Applications*, 388(23):4901–4909, 2009. ISSN 0378-4371. doi: <https://doi.org/10.1016/j.physa.2009.07.047>. URL <https://www.sciencedirect.com/science/article/pii/S0378437109006050>.
- [143] Q. Vera Liao, Werner Geyer, Michael Muller, and Yasaman Khazaen. *Conversational Interfaces for Information Search*. Springer International Publishing, Cham, 2020. ISBN 978-3-030-38825-6. doi: 10.1007/978-3-030-38825-6_13. URL https://doi.org/10.1007/978-3-030-38825-6_13.

- [144] Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. Pyserini: A python toolkit for reproducible information retrieval research with sparse and dense representations. In Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai, editors, *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 2356–2362. ACM, 2021. doi: 10.1145/3404835.3463238. URL <https://doi.org/10.1145/3404835.3463238>.
- [145] Jimmy Lin, Rodrigo Frassetto Nogueira, and Andrew Yates. *Pretrained Transformers for Text Ranking: BERT and Beyond*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2021. ISBN 978-3-031-01053-8. doi: 10.2200/S01123ED1V01Y202108HLT053. URL <https://doi.org/10.2200/S01123ED1V01Y202108HLT053>.
- [146] Lucy Linder, Michael Jungo, Jean Hennebert, Claudiu Cristian Musat, and Andreas Fischer. Automatic creation of text corpora for low-resource languages from the internet: The case of swiss german. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 2706–2711. European Language Resources Association, 2020. URL <https://aclanthology.org/2020.lrec-1.329/>.
- [147] Tie-Yan Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009. doi: 10.1561/15000000016. URL <https://doi.org/10.1561/15000000016>.
- [148] Rachel Tsz-Wai Lo, Ben He, and Iadh Ounis. Automatically building a stopword list for an information retrieval system. *Journal of Digital Information Management*, 3(1):3–8, 2005. URL <http://www.dirf.org/jdim/abstractv3i1.htm#01>.
- [149] Antoine Louis, Vageesh Kumar Saxena, Gijs van Dijck, and Gerasimos Spanakis. Colbert-xm: A modular multi-vector representation model for zero-shot multilingual information retrieval. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 4370–4383. Association for Computational Linguistics, 2025. URL <https://aclanthology.org/2025.coling-main.295/>.
- [150] Julie Beth Lovins. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11(1-2):22–31, 1968. URL <http://www.mt-archive.info/MT-1968-Lovins.pdf>.
- [151] Hans Peter Luhn. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1(4):309–317, 1957. doi: 10.1147/RD.14.0309. URL <https://doi.org/10.1147/rd.14.0309>.
- [152] Shengjie Ma, Chong Chen, Qi Chu, and Jiaxin Mao. Leveraging large language models for relevance judgments in legal case retrieval. *CoRR*, abs/2403.18405, 2024. doi: 10.48550/ARXIV.2403.18405. URL <https://doi.org/10.48550/arXiv.2403.18405>.

- [153] Xueguang Ma, Kai Sun, Ronak Pradeep, Minghan Li, and Jimmy Lin. Another look at DPR: reproduction of training and replication of retrieval. In Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørkvåg, and Vinay Setty, editors, *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part I*, volume 13185 of *Lecture Notes in Computer Science*, pages 613–626. Springer, 2022. doi: 10.1007/978-3-030-99736-6_41. URL https://doi.org/10.1007/978-3-030-99736-6_41.
- [154] Craig Macdonald and Nicola Tonellotto. Declarative experimentation in information retrieval using pyterrier. In Krisztian Balog, Vinay Setty, Christina Lioma, Yiqun Liu, Min Zhang, and Klaus Berberich, editors, *ICTIR '20: The 2020 ACM SIGIR International Conference on the Theory of Information Retrieval, Virtual Event, Norway, September 14-17, 2020*, pages 161–168. ACM, 2020. doi: 10.1145/3409256.3409829. URL <https://doi.org/10.1145/3409256.3409829>.
- [155] David J. C. MacKay and Linda C. Bauman Peto. A hierarchical dirichlet language model. *Nat. Lang. Eng.*, 1(3):289–308, 1995. doi: 10.1017/S1351324900000218. URL <https://doi.org/10.1017/S1351324900000218>.
- [156] Alexandre Magueresse, Vincent Carles, and Evan Heetderks. Low-resource languages: A review of past work and future challenges. *CoRR*, abs/2006.07264, 2020. URL <https://arxiv.org/abs/2006.07264>.
- [157] Thomas Mandl, Maristella Agosti, Giorgio Maria Di Nunzio, Alexander S. Yeh, Inderjeet Mani, Christine Doran, and Julia Maria Schulz. Logclef 2009: The CLEF 2009 multilingual logfile analysis track overview. In Carol Peters, Giorgio Maria Di Nunzio, Mikko Kurimo, Thomas Mandl, Djamel Mostefa, Anselmo Peñas, and Giovanna Roda, editors, *Multilingual Information Access Evaluation I. Text Retrieval Experiments, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 - October 2, 2009, Revised Selected Papers*, volume 6241 of *Lecture Notes in Computer Science*, pages 508–517. Springer, 2009. doi: 10.1007/978-3-642-15754-7_62. URL https://doi.org/10.1007/978-3-642-15754-7_62.
- [158] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *An Introduction to Information Retrieval*. Cambridge University Press, Cambridge, England, 2009. URL <https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf>.
- [159] Meta. Introducing meta llama 3: The most capable openly available llm to date, 2024. URL <https://llama.meta.com/llama3/>.
- [160] Sabrina J. Mielke, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y. Lee, Benoît Sagot, and Samson Tan. Between words and characters: A brief history of open-vocabulary modeling and tokenization in NLP. *CoRR*, abs/2112.10508, 2021. URL <https://arxiv.org/abs/2112.10508>.
- [161] Bhaskar Mitra and Nick Craswell. An introduction to neural information retrieval. *Foundations and Trends in Information Retrieval*, 13(1):1–126, 2018. doi: 10.1561/15000000061. URL <https://doi.org/10.1561/15000000061>.
- [162] Gordon Mohr, Michael Stack, Igor Ranitovic, Dan Avery, and Michele Kimpton. An introduction to heritrix - an open source archival quality web crawler. In *In IAWW'04, 4th*

- International Web Archiving Workshop*. Springer Press, 2004. URL <https://aclanthology.org/2022.dclrl-1.5.pdf>.
- [163] Marc Najork. Web crawler architecture. In Ling Liu and M. Tamer Özsu, editors, *Encyclopedia of Database Systems, Second Edition*. Springer, 2018. doi: 10.1007/978-1-4614-8265-9_457. URL https://doi.org/10.1007/978-1-4614-8265-9_457.
- [164] Flavio Cardoso Neves and Michael Mwara. Internet & intranet services in timor-leste. Technical report, Ministry of Transport and Communications, República Democrática de Timor-Leste, Dili, Timor-Leste, April 2006. URL <https://nsrc.org/regions/ASIA/TL/2006-Apr-11-East-Timor-and-the-Internet.pdf>.
- [165] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated machine reading comprehension dataset. In Tarek Richard Besold, Antoine Bordes, Artur S. d’Avila Garcez, and Greg Wayne, editors, *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2016. URL https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf.
- [166] Rubungo Andre Niyongabo, Hong Qu, Julia Kreutzer, and Li Huang. KINNEWS and KIRNEWS: benchmarking cross-lingual text classification for kinyarwanda and kirundi. In Donia Scott, Núria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8–13, 2020*, pages 5507–5521. International Committee on Computational Linguistics, 2020. doi: 10.18653/V1/2020.COLING-MAIN.480. URL <https://doi.org/10.18653/v1/2020.coling-main.480>.
- [167] Rodrigo Frassetto Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. Document ranking with a pretrained sequence-to-sequence model. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 708–718. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.FINDINGS-EMNLP.63. URL <https://doi.org/10.18653/v1/2020.findings-emnlp.63>.
- [168] National Institute of Linguistics (INL). *The Standard Orthography of the Tetum Language: 115 Years in the Making*. Directorate of the National Institute of Linguistics, Dili, Timor-Leste, August 2004. URL <https://archive.org/details/the-standard-orthography-of-the-tetum-language>.
- [169] Democratic Republic of Timor Leste. The standard orthography of the Tetun language. Government Decree-Law No. 1/2004 of 14 April 2024, 2004. URL <http://mj.gov.tl/jornal/lawsTL/RDTL-Law/RDTL-Gov-Decrees/Gov-Decree-2004-01.pdf>. Accessed on November 8, 2024.
- [170] The Government of Timor-Leste. Tatoli completes four years of existence, 2020. URL <http://timor-leste.gov.tl/?p=25225&lang=en&n=1>. Accessed on January 5, 2024.

- [171] The Foundation of Wikimedia. Wikimedia downloads, 2023. URL <https://dumps.wikimedia.org>. Accessed on February 19, 2024.
- [172] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. doi: 10.48550/ARXIV.2303.08774. URL <https://doi.org/10.48550/arXiv.2303.08774>.
- [173] Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald, and Douglas Johnson. Terrier information retrieval platform. In David E. Losada and Juan M. Fernández-Luna, editors, *Advances in Information Retrieval, 27th European Conference on IR Research, ECIR 2005, Santiago de Compostela, Spain, March 21-23, 2005, Proceedings*, volume 3408 of *Lecture Notes in Computer Science*, pages 517–519. Springer, 2005. doi: 10.1007/978-3-540-31865-1_37. URL https://doi.org/10.1007/978-3-540-31865-1_37.
- [174] Chris D. Paice. An evaluation method for stemming algorithms. In W. Bruce Croft and C. J. van Rijsbergen, editors, *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland, 3-6 July 1994 (Special Issue of the SIGIR Forum)*, pages 42–50. ACM/Springer, 1994. doi: 10.1007/978-1-4471-2099-5_5. URL https://doi.org/10.1007/978-1-4471-2099-5_5.
- [175] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL <https://aclanthology.org/P02-1040>.
- [176] Vassilis Plachouras, Ben He, and Iadh Ounis. University of glasgow at TREC 2004: Experiments in web, robust, and terabyte tracks with terrier. In Ellen M. Voorhees and Lori P. Buckland, editors, *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004*, volume 500-261 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2004. URL <http://trec.nist.gov/pubs/trec13/papers/uglasgow.web.robust.tera.pdf>.
- [177] Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In W. Bruce Croft, Alistair Moffat, C. J. van Rijsbergen, Ross Wilkinson, and Justin Zobel, editors, *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28 1998, Melbourne, Australia*, pages 275–281. ACM, 1998. doi: 10.1145/290941.291008. URL <https://doi.org/10.1145/290941.291008>.
- [178] Martin F. Porter. An algorithm for suffix stripping. *Program: electronic library and information systems*, 14(3):130–137, 1980. doi: 10.1108/EB046814. URL <https://doi.org/10.1108/eb046814>.
- [179] Rebecca Posner and Marius Sala. Occitan language. Encyclopedia Britannica, 2024. URL <https://www.britannica.com/topic/Occitan-language>. Accessed on February 19, 2024.
- [180] Tao Qin, Tie-Yan Liu, Jun Xu, and Hang Li. LETOR: A benchmark collection for research on learning to rank for information retrieval. *Inf. Retr.*, 13(4):346–374, 2010. doi: 10.1007/S10791-009-9123-Y. URL <https://doi.org/10.1007/s10791-009-9123-y>.

- [181] Chen Qu, Liu Yang, W. Bruce Croft, Johanne R. Trippas, Yongfeng Zhang, and Minghui Qiu. Analyzing and characterizing user intent in information-seeking conversations. In Kevyn Collins-Thompson, Qiaozhu Mei, Brian D. Davison, Yiqun Liu, and Emine Yilmaz, editors, *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 989–992. ACM, 2018. doi: 10.1145/3209978.3210124. URL <https://doi.org/10.1145/3209978.3210124>.
- [182] Filip Radlinski and Nick Craswell. A theoretical framework for conversational search. In Ragnar Nordlie, Nils Pharo, Luanne Freund, Birger Larsen, and Dan Russel, editors, *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval, CHIIR 2017, Oslo, Norway, March 7-11, 2017*, pages 117–126. ACM, 2017. doi: 10.1145/3020165.3020183. URL <https://doi.org/10.1145/3020165.3020183>.
- [183] Filip Radlinski, Madhu Kurup, and Thorsten Joachims. How does clickthrough data reflect retrieval quality? In James G. Shanahan, Sihem Amer-Yahia, Ioana Manolescu, Yi Zhang, David A. Evans, Aleksander Kolcz, Key-Sun Choi, and Abdur Chowdhury, editors, *Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, Napa Valley, California, USA, October 26-30, 2008*, pages 43–52. ACM, 2008. doi: 10.1145/1458082.1458092. URL <https://doi.org/10.1145/1458082.1458092>.
- [184] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020. URL <https://jmlr.org/papers/v21/20-074.html>.
- [185] Ruiyang Ren, Yingqi Qu, Jing Liu, Xin Zhao, Qifei Wu, Yuchen Ding, Hua Wu, Haifeng Wang, and Ji-Rong Wen. A thorough examination on zero-shot dense retrieval. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 15783–15796. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-EMNLP.1057. URL <https://doi.org/10.18653/v1/2023.findings-emnlp.1057>.
- [186] Stephen Robertson and Djoerd Hiemstra. Language models and probability of relevance. In J. Callan, B. Croft, and J. Lafferty, editors, *Proceedings of the Workshop on Language Modeling and Information Retrieval*, pages 21–25, United States, 2001. Carnegie Mellon University.
- [187] Stephen E. Robertson and Steve Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In W. Bruce Croft and C. J. van Rijsbergen, editors, *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland, 3-6 July 1994 (Special Issue of the SIGIR Forum)*, pages 232–241. ACM/Springer, 1994. doi: 10.1007/978-1-4471-2099-5_24. URL https://doi.org/10.1007/978-1-4471-2099-5_24.
- [188] Stephen E. Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/15000000019. URL <https://doi.org/10.1561/15000000019>.
- [189] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at TREC-3. In Donna K. Harman, editor, *Proceedings of The Third Text*

- REtrieval Conference, TREC 1994, Gaithersburg, Maryland, USA, November 2-4, 1994*, volume 500-225 of *NIST Special Publication*, pages 109–126. National Institute of Standards and Technology (NIST), 1994. URL <http://trec.nist.gov/pubs/trec3/papers/city.ps.gz>.
- [190] Shaurya Rohatgi, C. Lee Giles, and Jian Wu. What were people searching for? A query log analysis of an academic search engine. In J. Stephen Downie, Dana McKay, Hussein Suleman, David M. Nichols, and Faryaneh Poursardar, editors, *ACM/IEEE Joint Conference on Digital Libraries, JCDL 2021, Champaign, IL, USA, September 27-30, 2021*, pages 342–343. IEEE, 2021. doi: 10.1109/JCDL52503.2021.00062. URL <https://doi.org/10.1109/JCDL52503.2021.00062>.
- [191] Siba Sankar Sahu and Sukomal Pal. Building a text retrieval system for the sanskrit language: Exploring indexing, stemming, and searching issues. *Computer Speech and Language*, 81:101518, 2023. doi: 10.1016/J.CSL.2023.101518. URL <https://doi.org/10.1016/j.csl.2023.101518>.
- [192] Siba Sankar Sahu, Debrup Dutta, Sukomal Pal, and Imran Rasheed. Effect of stopwords and stemming techniques in urdu IR. *SN Computer Science*, 4(5):547, 2023. doi: 10.1007/S42979-023-01953-4. URL <https://doi.org/10.1007/s42979-023-01953-4>.
- [193] Shadi Saleh and Pavel Pecina. CUNI at the ShARe/CLEF eHealth evaluation lab 2014. In Linda Cappellato, Nicola Ferro, Martin Halvey, and Wessel Kraaij, editors, *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014*, volume 1180 of *CEUR Workshop Proceedings*, pages 226–235. CEUR-WS.org, 2014. URL <https://ceur-ws.org/Vol-1180/CLEF2014wn-eHealth-SalehEt2014.pdf>.
- [194] Mark Sanderson. Test collection based evaluation of information retrieval systems. *Foundations and Trends in Information Retrieval*, 4(4):247–375, 2010. doi: 10.1561/1500000009. URL <https://doi.org/10.1561/1500000009>.
- [195] Mark Sanderson and W. Bruce Croft. The history of information retrieval research. *Proc. IEEE*, 100(Centennial-Issue):1444–1451, 2012. doi: 10.1109/JPROC.2012.2189916. URL <https://doi.org/10.1109/JPROC.2012.2189916>.
- [196] Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. Colbertv2: Effective and efficient retrieval via lightweight late interaction. In Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 3715–3734. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.NAACL-MAIN.272. URL <https://doi.org/10.18653/v1/2022.naacl-main.272>.
- [197] Tefko Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. part III: behavior and effects of relevance. *Journal of the American Society for Information Science and Technology*, 58(13):2126–2144, 2007. doi: 10.1002/ASI.20681. URL <https://doi.org/10.1002/asi.20681>.
- [198] Shota Sasaki, Shuo Sun, Shigehiko Schamoni, Kevin Duh, and Kentaro Inui. Cross-lingual learning-to-rank with shared representations. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the*

- Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 458–463. Association for Computational Linguistics, 2018. doi: 10.18653/V1/N18-2073. URL <https://doi.org/10.18653/v1/n18-2073>.
- [199] Jacques Savoy. A stemming procedure and stopword list for general french corpora. *Journal of the American Society for Information Science*, 50(10):944–952, 1999. doi: 10.5555/318976.318984. URL <https://dl.acm.org/doi/10.5555/318976.318984>.
- [200] Kevin P. Scannell. The crúbadán project: Corpus building for under-resourced languages. In *Proceedings of the 3rd Web as Corpus Workshop (WAC3)*, pages 5–15, Louvain-la-Neuve, Belgium, 2007.
- [201] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x. URL <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>.
- [202] Craig Silverstein, Monika Henzinger, Hannes Marais, and Michael Moricz. Analysis of a very large web search engine query log. *SIGIR Forum*, 33(1):6–12, 1999. doi: 10.1145/331403.331405. URL <https://doi.org/10.1145/331403.331405>.
- [203] Rion Snow, Brendan O’Connor, Daniel Jurafsky, and Andrew Y. Ng. Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks. In *2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu, Hawaii, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 254–263. ACL, 2008. URL <https://aclanthology.org/D08-1027/>.
- [204] Snowball. Stemming algorithms for use in information retrieval, 2002. URL <https://snowballstem.org>. Accessed on February 19, 2024.
- [205] Snowball. Spanish stemming algorithm, 2005. URL <https://snowballstem.org/algorithms/spanish/stemmer.html>.
- [206] Snowball. Portuguese stemming algorithm, 2005. URL <https://snowballstem.org/algorithms/portuguese/stemmer.html>.
- [207] Ian Soboroff, Shudong Huang, and Donna Harman. TREC 2020 news track overview. In Ellen M. Voorhees and Angela Ellis, editors, *Proceedings of the Twenty-Ninth Text REtrieval Conference, TREC 2020, Virtual Event [Gaithersburg, Maryland, USA], November 16-20, 2020*, volume 1266 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2020. URL <https://trec.nist.gov/pubs/trec29/papers/OVERVIEW.N.pdf>.
- [208] Eero Sormunen. Liberal relevance criteria of TREC -: counting on negligible documents? In Kalervo Järvelin, Micheline Beaulieu, Ricardo A. Baeza-Yates, and Sung-Hyon Myaeng, editors, *SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 11-15, 2002, Tampere, Finland*, pages 324–330. ACM, 2002. doi: 10.1145/564376.564433. URL <https://doi.org/10.1145/564376.564433>.

- [209] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 1972. doi: 10.1108/eb026526. URL <https://doi.org/10.1108/eb026526>.
- [210] Karen Spärck-Jones and Cornelis Joost van Rijsbergen. Report on the need for and the provision of an ‘ideal’ information retrieval test collection. Technical Report British Library Research and Development Report No. 5266, Computer Laboratory, University of Cambridge, Cambridge, United Kingdom, 1975. URL https://sigir.org/files/museum/pub-14/pub_14.pdf.
- [211] Stephanie M. Strassel and Jennifer Tracey. LORELEI language packs: Data, tools, and resources for technology development in low resource languages. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA), 2016. URL <http://www.lrec-conf.org/proceedings/lrec2016/summaries/1138.html>.
- [212] Hussein Suleman. Information retrieval in african languages. *CoRR*, abs/1806.04735, 2018. URL <http://arxiv.org/abs/1806.04735>.
- [213] R. M. Suresh and R. Padmajavalli. An overview of data preprocessing in data and web usage mining. In *First IEEE International Conference on Digital Information Management (ICDIM), December 6-8, 2006, Christ College, Bangalore, India, Proceedings*, pages 193–198. IEEE, 2006. doi: 10.1109/ICDIM.2007.369352. URL <https://doi.org/10.1109/ICDIM.2007.369352>.
- [214] About Switzerland. Language – facts and figures, 2024. URL <https://www.eda.admin.ch/aboutswitzerland/en/home/gesellschaft/sprachen/die-sprache-n---fakten-und-zahlen.html>. Accessed on February 19, 2024.
- [215] Bilal Tahir and Muhammad Amir Mehmood. Corpulyzer: A novel framework for building low resource language corpora. *IEEE Access*, 9:8546–8563, 2021. doi: 10.1109/ACCESS.2021.3049793. URL <https://doi.org/10.1109/ACCESS.2021.3049793>.
- [216] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In Joaquin Vanschoren and Sai-Kit Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021. URL <https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/65b9eea6e1cc6bb9f0cd2a47751a186f-Abstract-round2.html>.
- [217] Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. Large language models can accurately predict searcher preferences. In Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang, editors, *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pages 1930–1940. ACM, 2024. doi: 10.1145/3626772.3657707. URL <https://doi.org/10.1145/3626772.3657707>.

- [218] Johanne R. Trippas, Sara Fahad Dawood Al Lawati, Joel Mackenzie, and Luke Gallagher. What do users really ask large language models? an initial log analysis of google bard interactions in the wild. In Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang, editors, *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pages 2703–2707. ACM, 2024. doi: 10.1145/3626772.3657914. URL <https://doi.org/10.1145/3626772.3657914>.
- [219] Sarah K. Tyler and Jaime Teevan. Large scale query log analysis of re-finding. In Brian D. Davison, Torsten Suel, Nick Craswell, and Bing Liu, editors, *Proceedings of the Third International Conference on Web Search and Web Data Mining, WSDM 2010, New York, NY, USA, February 4-6, 2010*, pages 191–200. ACM, 2010. doi: 10.1145/1718487.1718512. URL <https://doi.org/10.1145/1718487.1718512>.
- [220] UNESCO. World atlas of languages, 2024. URL <https://en.wal.unesco.org/languages/mizo>. Accessed on February 19, 2024.
- [221] UMINHO University of Minho. words-pt: Dicionário natura, 2021. URL <https://natura.di.uminho.pt/download/sources/Dictionaries/wordlists/>. Accessed on May 1, 2024.
- [222] Suzanne Urbanczyk. Phonological and morphological aspects of reduplication. Oxford University Press, 03 2017. URL <https://doi.org/10.1093/acrefore/9780199384655.013.80>.
- [223] Catharina Williams van Klinken and John Hajek. Language contact and functional expansion in Tetun Dili: The evolution of a new press register. *Multilingua*, 37:613–647, 2018. URL <https://doi.org/10.1515/multi-2017-0109>.
- [224] Catharina Williams van Klinken, John Hajek, and Rachel Nordlingerh. *Tetun Dili: a grammar of an East Timorese language*. Pacific Linguistics, Canberra, Australia, 2002. ISBN 85883-509-6. doi: 10.15144/PL-528. URL <http://hdl.handle.net/1885/146149>.
- [225] Artur Varanda, Leonel Santos, Rogério Luís de Carvalho Costa, Adail Oliveira, and Carlos Rabadão. The general data protection regulation and log pseudonymization. In Leonard Barolli, Isaac Woungang, and Tomoya Enokido, editors, *Advanced Information Networking and Applications - Proceedings of the 35th International Conference on Advanced Information Networking and Applications (AINA-2021), Toronto, ON, Canada, 12-14 May, 2021, Volume 3*, volume 227 of *Lecture Notes in Networks and Systems*, pages 479–490. Springer, 2021. doi: 10.1007/978-3-030-75078-7_48. URL https://doi.org/10.1007/978-3-030-75078-7_48.
- [226] 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. Constituição anotada da República Democrática de Timor-Leste. Escola de Direito da Universidade do Minho, 2011. URL <http://hdl.handle.net/10400.22/4008>.
- [227] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N.

- Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. URL <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>.
- [228] Andreas von Holy, Alon Bresler, Osher Shuman, Catherine Chavula, and Hussein Suleman. Bantuweb: a digital library for resource scarce south african languages. In Muthoni Masinde, editor, *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, 2017. doi: 10.1145/3129416.3129446. URL <https://doi.org/10.1145/3129416.3129446>.
- [229] Ellen M. Voorhees. Overview of TREC 2004. In Ellen M. Voorhees and Lori P. Buckland, editors, *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004*, volume 500-261 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2004. URL <http://trec.nist.gov/pubs/trec13/papers/OVERVIEW13.pdf>.
- [230] Ellen M. Voorhees. Overview of the TREC 2004 robust track. In Ellen M. Voorhees and Lori P. Buckland, editors, *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004, Gaithersburg, Maryland, USA, November 16-19, 2004*, volume 500-261 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2004. URL <http://trec.nist.gov/pubs/trec13/papers/ROBUST.OVERVIEW.pdf>.
- [231] Ellen M. Voorhees. Overview of the TREC 2006. In Ellen M. Voorhees and Lori P. Buckland, editors, *Proceedings of the Fifteenth Text REtrieval Conference, TREC 2006, Gaithersburg, Maryland, USA, November 14-17, 2006*, volume 500-272 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2006. URL <http://trec.nist.gov/pubs/trec15/papers/OVERVIEW.pdf>.
- [232] Ellen M. Voorhees and Donna Harman. The Text REtrieval Conferences (TREC). In *TIPSTER TEXT PROGRAM PHASE III: Proceedings of a Workshop held at Baltimore, Maryland, October 13-15, 1998*, pages 241–273, Baltimore, Maryland, USA, October 1998. Association for Computational Linguistics. doi: 10.3115/1119089.1119127. URL <https://aclanthology.org/X98-1031>.
- [233] Ellen M. Voorhees and Donna Harman. Overview of the ninth text retrieval conference (TREC-9). In Ellen M. Voorhees and Donna K. Harman, editors, *Proceedings of The Ninth Text REtrieval Conference, TREC 2000, Gaithersburg, Maryland, USA, November 13-16, 2000*, volume 500-249 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2000. URL http://trec.nist.gov/pubs/trec9/papers/overview_9.pdf.
- [234] Jiajia Wang, Jimmy Xiangji Huang, Xinhui Tu, Junmei Wang, Angela Jennifer Huang, Md. Tahmid Rahman Laskar, and Amran Bhuiyan. Utilizing BERT for information retrieval: Survey, applications, resources, and challenges. *ACM Comput. Surv.*, 56(7):185:1–185:33, 2024. doi: 10.1145/3648471. URL <https://doi.org/10.1145/3648471>.
- [235] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. Ccnet: Extracting high quality monolingual datasets from web crawl data. In Nicoletta Calzolari, Frédéric Béchet, Philippe

- Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 4003–4012. European Language Resources Association, 2020. URL <https://aclanthology.org/2020.lrec-1.494/>.
- [236] Matt Wiens. english-words-py. PyPI Package, 2021. URL <https://pypi.org/project/english-words/>. Accessed on May 1, 2024.
- [237] Papers with Code. Multi-task language understanding on mmlu, 2024. URL <https://paperswithcode.com/sota/multi-task-language-understanding-on-mmlu>. Accessed on April 20, 2024.
- [238] Elisabeth Wolf, Delphine Bernhard, and Iryna Gurevych. Combining probabilistic and translation-based models for information retrieval based on word sense annotations. In Carol Peters, Giorgio Maria Di Nunzio, Mikko Kurimo, Thomas Mandl, Djamel Mostefa, Anselmo Peñas, and Giovanna Roda, editors, *Multilingual Information Access Evaluation I. Text Retrieval Experiments, 10th Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece, September 30 - October 2, 2009, Revised Selected Papers*, volume 6241 of *Lecture Notes in Computer Science*, pages 120–127. Springer, 2009. doi: 10.1007/978-3-642-15754-7_14. URL https://doi.org/10.1007/978-3-642-15754-7_14.
- [239] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.EMNLP-DEMOS.6. URL <https://doi.org/10.18653/v1/2020.emnlp-demos.6>.
- [240] Hao Wu and Hui Fang. Tie breaker: A novel way of combining retrieval signals. In Oren Kurland, Donald Metzler, Christina Lioma, Birger Larsen, and Peter Ingwersen, editors, *International Conference on the Theory of Information Retrieval, ICTIR '13, Copenhagen, Denmark, September 29 - October 02, 2013*, page 16. ACM, 2013. doi: 10.1145/2499178.2499192. URL <https://doi.org/10.1145/2499178.2499192>.
- [241] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL <https://openreview.net/forum?id=zeFrfgYzln>.
- [242] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou,

- editors, *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 483–498. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.NAACL-MAIN.41. URL <https://doi.org/10.18653/v1/2021.naacl-main.41>.
- [243] Wei Yang, Haotian Zhang, and Jimmy Lin. Simple applications of BERT for ad hoc document retrieval. *CoRR*, abs/1903.10972, 2019. URL <http://arxiv.org/abs/1903.10972>.
- [244] Xinyan Yu, Trina Chatterjee, Akari Asai, Junjie Hu, and Eunsol Choi. Beyond counting datasets: A survey of multilingual dataset construction and necessary resources. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *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, 2022. doi: 10.18653/V1/2022.FINDINGS-EMNLP.273. URL <https://doi.org/10.18653/v1/2022.findings-emnlp.273>.
- [245] Hamed Zamani, Johanne R. Trippas, Jeff Dalton, and Filip Radlinski. Conversational information seeking. *Found. Trends Inf. Retr.*, 17(3-4):244–456, 2023. doi: 10.1561/15000000081. URL <https://doi.org/10.1561/15000000081>.
- [246] Marcos Zampieri. 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, 2013. doi: 10.1109/CINTI.2013.6705230.
- [247] ChengXiang Zhai and John D. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In W. Bruce Croft, David J. Harper, Donald H. Kraft, and Justin Zobel, editors, *SIGIR 2001: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, September 9-13, 2001, New Orleans, Louisiana, USA*, pages 334–342. ACM, 2001. doi: 10.1145/383952.384019. URL <https://doi.org/10.1145/383952.384019>.
- [248] Jie Zhang and Ali A. Ghorbani. The reconstruction of user sessions from a server log using improved time-oriented heuristics. In *2nd Annual Conference on Communication Networks and Services Research (CNSR 2004), 19-21 May 2004, Fredericton, N.B., Canada*, pages 315–322. IEEE Computer Society, 2004. doi: 10.1109/DNSR.2004.1344744. URL <https://doi.org/10.1109/DNSR.2004.1344744>.
- [249] Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. Mr. TyDi: A multi-lingual benchmark for dense retrieval. In Duygu Ataman, Alexandra Birch, Alexis Conneau, Orhan Firat, Sebastian Ruder, and Gozde Gul Sahin, editors, *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 127–137, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.mrl-1.12. URL <https://aclanthology.org/2021.mrl-1.12/>.
- [250] Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. MIRACL: A multilingual retrieval dataset covering 18 diverse languages. *Trans. Assoc. Comput. Linguistics*, 11:1114–1131, 2023. doi: 10.1162/TACL_A_00595. URL https://doi.org/10.1162/tacl_a_00595.

- [251] Xinyu Zhang, Kelechi Ogueji, Xueguang Ma, and Jimmy Lin. Toward best practices for training multilingual dense retrieval models. *ACM Trans. Inf. Syst.*, 42(2):39:1–39:33, 2024. doi: 10.1145/3613447. URL <https://doi.org/10.1145/3613447>.
- [252] Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. Dense text retrieval based on pretrained language models: A survey. *ACM Trans. Inf. Syst.*, 42(4):89:1–89:60, 2024. doi: 10.1145/3637870. URL <https://doi.org/10.1145/3637870>.
- [253] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P. Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. Lmsys-chat-1m: A large-scale real-world LLM conversation dataset. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL <https://openreview.net/forum?id=BOfDKxfwt0>.
- [254] Mu Zhu. Recall, precision and average precision. *Department of Statistics and Actuarial Science, University of Waterloo, Waterloo*, 2(30):6, 2004. URL <https://datascience-intro.github.io/1MS041-2022/Files/AveragePrecision.pdf>.
- [255] Shengyao Zhuang, Xueguang Ma, Bevan Koopman, Jimmy Lin, and Guido Zuccon. Promptreps: Prompting large language models to generate dense and sparse representations for zero-shot document retrieval. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 4375–4391. Association for Computational Linguistics, 2024. URL <https://aclanthology.org/2024.emnlp-main.250>.
- [256] George Kingsley Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley Press, 1949. URL https://www.iqila.org/includes/basic_references/Zipf_1949_Human_Behavior_and_the_Principle_of_Least_Effort.pdf.
- [257] Justin Zobel. What we talk about when we talk about information retrieval. *SIGIR Forum*, 51(3):18–26, 2017. doi: 10.1145/3190580.3190584. URL <https://doi.org/10.1145/3190580.3190584>.