SEA-BED: Southeast Asia Embedding Benchmark

Wuttikorn Ponwitayarat[1*], Raymond Ng[2*], Jann Railey Montalan[2], Thura Aung[5], Jian Gang Ngui[2], Yosephine Susanto[2], William Tjhi[2], Panuthep Tasawong[1], Erik Cambria[3], Ekapol Chuangsuwanich[4], Sarana Nutanong[1], Peerat Limkonchotiwat[2*]

[1] Vidyasirimedhi Institute of Science and Technology
wuttikorn.p_s22@vistec.ac.th, panuthep.t_s20@vistec.ac.th, snutanon@vistec.ac.th

[2] AI Singapore
raymond@aisingapore.org, railey@aisingapore.org, jiangangngui@aisingapore.org, yosephine@aisingapore.org, wtjhi@aisingapore.org, peerat@aisingapore.org

[3] Nanyang Technological University
cambria@ntu.edu.sg

[4] Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University
ekapolc@cp.eng.chula.ac.th

[5] King Mongkut's Institute of Technology Ladkrabang
66011606@kmitl.ac.th

*Sentence embeddings are essential for NLP tasks such as semantic search, re-ranking, and textual similarity. Although multilingual benchmarks like MMTEB broaden coverage, **S**outh**e**ast **A**sia (SEA) datasets are scarce and often machine-translated, missing native linguistic properties. With nearly 700 million speakers, the SEA region lacks a region-specific embedding benchmark. We introduce **SEA-BED**, the first large-scale SEA embedding benchmark with 169 datasets across 9 tasks and 10 languages, where 71% are formulated by humans, not machine generation or translation. We address three research questions: (1) which SEA languages and tasks are challenging, (2) whether SEA languages show unique performance gaps globally, and (3) how human vs. machine translations affect evaluation. We evaluate 17 embedding models across six studies, analyzing task and language challenges, cross-benchmark comparisons, and translation trade-offs. Results show sharp ranking shifts, inconsistent model performance among SEA languages, and the importance of human-curated datasets for low-resource languages like Burmese.* [1]

0 *Equal contribution

1 Link to SEA-BED: [HIDDEN] (we will add the link to our benchmark and codes in the published version of our paper)

1. Introduction

Sentence embedding plays a crucial role in Natural Language Processing (NLP) by transforming complex linguistic structures into fixed-size vector representations that capture semantic meaning. These embeddings are fundamental for various downstream tasks, including semantic textual similarity, retrieval, and re-ranking. To evaluate the robustness of embeddings, numerous studies employ benchmark suites covering these tasks (Muennighoff et al. 2023; Enevoldsen et al. 2025). However, most existing benchmarks, such as SentEval (Conneau and Kiela 2018) and MTEB (Muennighoff et al. 2023), mainly focus on high-resource languages like English, German, Chinese, and French, leading to the underrepresentation of low-resource languages.

Recently, several works have proposed multilingual sentence embedding models and benchmarks. For example, Wang et al. (2024a) built a model based on Mistral-7B (Jiang et al. 2023) supporting 93 languages and Jina-embedding-v3 (Sturua et al. 2024a) covering 89 languages spoken worldwide. To broaden benchmarking efforts, Enevoldsen et al. (2025) proposed MMTEB, an extension of MTEB that includes evaluation across 1,090 languages. Despite these advances and coverage, **S**outh**e**ast **A**sian (SEA) languages remain underrepresented. Existing SEA resources in MMTEB–XNLI (Conneau et al. 2018), Tatoeba (community 2021), and SIB-200 (Adelani et al. 2024)–are machine-translated datasets, typically translated from English sentences to SEA languages. This lack of native-authored data limits the fluency and linguistic authenticity captured in current benchmarks. As a result, sentence embedding models may struggle to generalize effectively for SEA languages, highlighting a critical gap in current multilingual sentence embedding research.

Even though the combined population of the SEA region is close to 700 million [2], and is the 3rd largest population group in the world, no SEA sentence embedding benchmarks have been established so far. Prior efforts, i.e., IndoNLG (Aji et al. 2022), SEACrowd (Lovenia et al. 2024), and SEA-VL (Cahyawijaya et al. 2025), collect more than 500 datasets and 100 million samples for SEA. Still, these resources were primarily designed for training and evaluating language generation models (i.e., decoder-based models) and are unsuitable for sentence embedding tasks. Moreover, there are many works studied on SEA languages via large language models (LLMs), i.e., SEA-LION (Ng et al. 2025) and SEA-LLMs (Zhang et al. 2024). These works underscore the importance of specifically designed benchmarks to identify the challenges and gaps in collecting and processing SEA language data. Building upon these insights, our work addresses the following research questions: **(RQ1)** Within SEA languages, which specific tasks, scripts, or linguistic features remain problematic for current models, and why do these remain unresolved? **(RQ2)** How do current multilingual embedding models perform in SEA languages compared to other global languages, and do we see unique performance gaps in SEA? **(RQ3)** What trade-offs arise between human-annotated versus machine-generated data?

To answer our research questions, we propose SEA-BED, a **SEA** sentence em**BED**ding benchmark that collects datasets created by native speakers rather than relying primarily on translation. Our benchmark contains more than 169 datasets, 9 tasks, and 10 languages. Crucially, 148 datasets (∼87% of our benchmark) never appear in any sentence embedding benchmarks, and 120 datasets (∼71% of our benchmark) are created by humans rather than machine translation or generation. This makes our benchmark representative of the SEA region more so than any other benchmark by evaluating sentence embedding models on human-crafted datasets, resulting in more reliable and accurate results than machine translation or generation datasets. Moreover, we propose 11 new datasets for Thai and Burmese, which allow us to evaluate the semantic textual similarity,

2 United Nations, Economic and Social Commission for Asia and the Pacific (ESCAP), 2024

relation understanding, and transfer learning performances. We summarize the differences between our benchmark and previous benchmarks in Table 1.

To evaluate the efficiency of sentence embeddings, we tested 17 embedding models in six distinct studies (Section 5) for both encoder- and decoder-based architectures that support SEA and other languages: **Sections 5.1 and 5.2** address RQ1 by breaking down model performance across tasks and languages within SEA. **Section 5.3** answers RQ2, comparing SEA-focused performance to the broader MMTEB setting. **Section 5.4** focuses on RQ3's design decision analysis, contrasting human-annotated data with machine-translated data in Thai and Burmese. **Section 5.5 and 5.6** dive deeper into RQ1 with studies to understand the challenge of our benchmark using a tokenizer-level analysis and a SEA language similarity study.

The experimental results from our studies show that the performance of existing multilingual sentence embedding significantly changes compared to MMTEB (Enevoldsen et al. 2025). For example, the ranking of the model that used to perform best over 1,090 languages dropped from second rank to eighth rank when tested on our benchmark, as shown in Figure 2. We propose that this change in ranking is due to our benchmark including more high-quality SEA datasets than MMTEB, which makes our benchmark more challenging, and crucially more holistically representative for the SEA region at large. Moreover, we found performance and language consistency problems, where no models perform best in all cases for SEA languages.

Summary of the contributions of our paper:

- We propose SEA-BED, a massive collection of high-quality SEA datasets. Our benchmark consists of more than 169 datasets, 9 tasks, and 10 languages: Indonesian, Thai, Vietnamese, Burmese, Filipino, Tamil, Khmer, Malay, Lao, and Tetum.
- We propose 11 new datasets for Thai and Burmese, allowing us to evaluate more tasks compared to previous benchmarks. For reproducibility, we will release the evaluation tool and datasets in the published version of our paper.
- We conduct a comprehensive empirical study on SEA languages by evaluating 17 embedding models across 6 distinct studies. Our experiments reveal that performance rankings for multilingual embeddings shift significantly when tested on SEA data, and no single model excels across all tasks and languages.

Benchmark	# Languages	# SEA Languages	# Datasets	# Task	# New datasets	# SEA datasets	# Human-Crafted datasets (only SEA languages)
MTEB-French (Ciancone et al. 2024a)	1	N/A	18	8	3	N/A	N/A
C-Pack (Xiao et al. 2024a)	1	N/A	35	6	35	N/A	N/A
SEB (Enevoldsen et al. 2024)	4	N/A	24	4	24	N/A	N/A
MMTEB (Enevoldsen et al. 2025)	1,090	9	270	10	5	21 (7.78 %)	20 (95.24 %)
SEA-BED (ours)	10	10	169	9	11	169 (100.00 %)	120 (71.01 %)

Table 1: The statistics of our benchmark compared to existing sentence embedding benchmarks.

2. Related Work

2.1 Embedding Models

Currently, researchers typically use pre-trained language models and contrastive learning to train text embedding models (Li and Li 2023; Wang et al. 2024a; Lee et al. 2024). Feng et al. (2022) proposed a language-agnostic multilingual sentence embedding on 109 languages called LaBSE. The experimental results from LaBSE demonstrate robust performance across cross-lingual and multilingual retrieval benchmarks. Wang et al. (2024b) proposed pre-training and fine-tuning methods where the pre-trained models are designed for specific sentence embedding tasks. In particular, the models are trained using unsupervised contrastive learning and then further fine-tuned with supervised contrastive learning to improve the robustness. Chen et al. (2024) proposed BGE-M3, an effective sentence embedding model that leverages a combination of sparse and

dense models in the training step. The experimental results of previous works demonstrated state-of-the-art performance in various languages, including SEA languages, such as Thai, Indonesian, and Filipino.

Recent advancements in LLMs have led to the development of multilingual text models. For example, Wang et al. (2024a) demonstrated that an LLM such as GPT-4 can synthetically generate pre-training and fine-tuning data for sentence embedding models. Moreover, Wang et al. (2024a) proposed e5-instruction, an LLM-based sentence embedding that pools the last layer of Mistral-7B (Jiang et al. 2023) to formulate the sentence representation of text. The instruction will affect the representation of texts in downstream tasks, e.g., the prompts of classification and relation tasks will give different representations despite identical text input. Muennighoff et al. (2024a) proposed a technique to combine the text representation and instruction following. In particular, the representation is learned through contrastive learning, while the instruction learning is incorporated during the training step using a Supervised Fine-Tuning (SFT) technique.

2.2 Text Embedding Benchmarks

Existing text embedding benchmarks primarily focus on high-resource languages. Notable examples include SentEval (Conneau and Kiela 2018), which provides a preliminary benchmark for text embedding understanding of STS and transfer learning. USEB (Wang, Reimers, and Gurevych 2021) is an unsupervised embedding benchmark focusing on pair-text classification, such as re-ranking, paraphrase detection, and information retrieval. BEIR (Thakur et al. 2021) is a Heterogeneous Benchmark focusing only on information retrieval datasets on 18 datasets. MTEB (Muennighoff et al. 2023) is a large-scale version of BEIR that not only focuses on retrieval tasks but also on diverse tasks, i.e., bitext mining, classification, and semantic textual similarity. However, these benchmarks primarily focus on English, while many works extend MTEB from English to Chinese (Xiao et al. 2024b), German (Wehrli, Arnrich, and Irrgang 2023), and French (Ciancone et al. 2024b).

Recently, there has been an attempt to create a multilingual version of MTEB called MMTEB (Enevoldsen et al. 2025). This recent multilingual benchmark evaluates 10 tasks and 1,090 datasets, but notably, only 21 datasets are SEA language datasets. The experiment results from MMTEB found that many multilingual embedding models that perform well in English might fail on this benchmark because they lack consistency in cross-lingual settings. However, we found that the SEA texts in MMTEB benchmarks are not written naturally due to the fact that the datasets are created through machine translation of English to SEA texts. Thus, the results from MMTEB might not be representative of performance in SEA languages, given the reliance on machine-translated datasets.

2.3 SEA Benchmarks

There have been many efforts to formulate SEA benchmarks. NusaCrowd (Cahyawijaya et al. 2023) proposed a large-scale Indonesian benchmark focusing on natural language understanding and generation, especially for decoder models. SEACrowd (Lovenia et al. 2024) is a data collection project that gathers SEA benchmarks in its own repository. The experiment from SEACrowd is focused only on large language models, especially Llama (Dubey et al. 2024) and T5 (Raffel et al. 2020) families. SEA-VQA (Urailertprasert et al. 2024) proposed a vision question-answering dataset where the image is SEA cultural collected from the UNESCO website, and the question and answer are written only in English. However, these benchmarks do not measure the embedding effectiveness in SEA texts. In particular, previous works studied large language models and generative outputs, while embeddings have not been experimented with in SEA languages.

Furthermore, there have also been many efforts to create SEA models that can perform well on SEA benchmarks. SEA-LION (Ng et al. 2025) and SeaLLMs (Zhang et al. 2024) proposed a Southeast Asian model using existing large language models (Dubey et al. 2024; Rivière et al. 2024) as their base models. The performance of those models on generative SEA benchmarks (Susanto et al. 2025) demonstrated the effectiveness of both models in terms of culture and knowledge of SEA texts. Nonetheless, the robustness of those models from the embedding perspective has yet to be explored.

3. SEA-BED

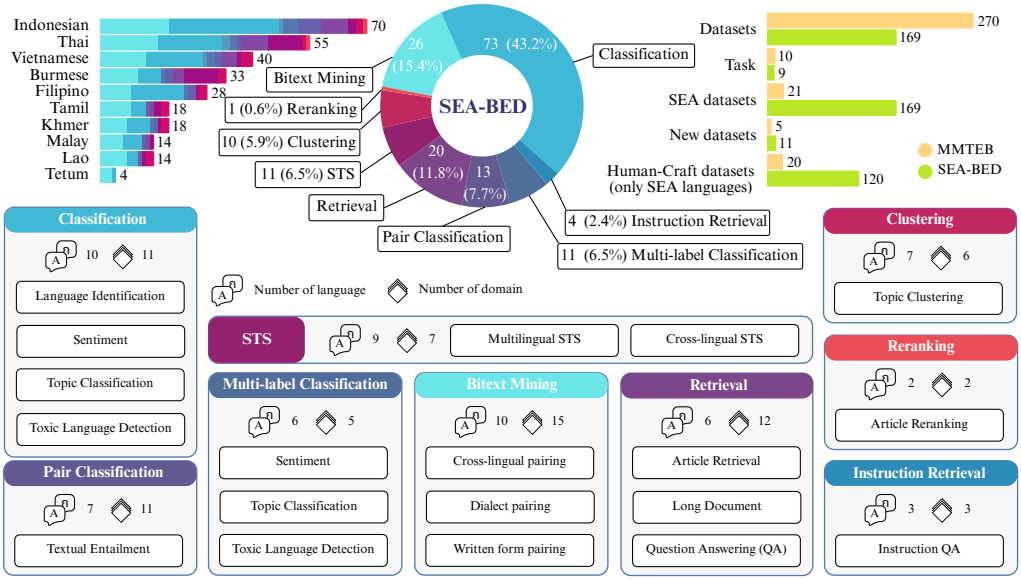


Figure 1: An overview of SEA-BED, featuring 169 datasets, 9 tasks, and 10 languages.

3.1 Overview

As shown in Figure 1, our benchmark consists of 169 open-source datasets written in 10 languages: Indonesian, Thai, Vietnamese, Burmese, Filipino, Tamil, Khmer, Malay, Lao, and Tetum. The goals of SEA-BED are:

- Evaluate the cross-lingual and multilingual capability of existing sentence embedding models for both open-source and proprietary models.
- Evaluate the robustness and consistency of sentence embedding models. The robust models should perform similarly regardless of tasks or input languages.
- Holistic results for SEA by analyzing more on SEA studies, including results on machine translation datasets, human-crafted datasets, tokenizers regarding SEA tokens, and language similarity.
- Open-data, open-result, and open-science. We will release the codes and datasets publicly for reproducibility.

In contrast to the previous sentence embedding benchmarks, as shown in Table 1, we consolidate the SEA datasets into our benchmark, resulting in an evaluation suite that better represents SEA knowledge than previous works. Note that the complete datasets and download links are provided in Appendix 1.2.

3.2 Task Selection and Evaluation

To evaluate SEA embedding in all aspects, we select diverse tasks based on real-world applications of embeddings. However, because SEA resources are limited, we select 9 tasks for which the resources are available and open-source, such as bitext mining, pair classification, classification, clustering, semantic textual similarity, retrieval, reranking, multi-label classification, and instruction retrieval.

We categorize the selected tasks into 4 aspects reflecting the desired properties that we want to evaluate:

- **Cross-lingual Evaluation (CLE).** In this task type, we evaluate the cross-lingual knowledge of embedding models where the inputs can be written in different languages in the same dataset.
- **Transfer Learning (TL).** For this task type, we evaluate the robustness of embeddings in downstream tasks by following SentEval (Conneau and Kiela 2018) and use a classifier with the embedding model where the classifier learns to classify the class of text (i.e., sentiment analysis, review classification, etc.) according to the target task.
- **Semantic Understanding (SU).** In this task type, we evaluate the robustness of producing similarity of text-pairs. A good model should generate a similar or dissimilar text-pair when the input is relevant or irrelevant, respectively.
- **Information Retrieval (IR).** For the final task type, we evaluate the pair-wise similarity generation. In particular, we observe the retrieval performance when the query and documents can be significantly different in length, unlike the **SU** category, in which the input lengths are always similar.

Along with the above-defined task types, we add a tag in the selected tasks below to easily understand *what desired property we want to evaluate for each task*.

[CLE, IR] Bitext Mining. Given two sentence sets from different languages, the goal is to find the best match for each sentence in the first set, typically its translation. The model embeds each sentence, and matches are determined using cosine similarity. We used the F1 score to evaluate model performance.

[SU] Pair Classification. This task involves a pair of input sentences, with their relationship indicated by a binary label. The relationship is predicted based on embedding similarity, using average precision as the primary metric.

[TL] Classification. The train and test sets are embedded using the provided model. A logistic regression classifier is applied on top of the embeddings, and the model's performance is evaluated on the test set using the F1 score.

[TL, SU] Clustering. Given a collection of sentences or paragraphs, the task is to organize them into meaningful clusters. The embedded texts are clustered using a mini-batch k-means algorithm, with the number of clusters (k) set to the number of unique labels. Clustering performance is evaluated using the V-measure metric.

[CLE, SU] Semantic Textual Similarity (STS). The STS task measures sentence similarity using continuous labels, with higher values indicating greater similarity. The model embeds a pair of sentences, and similarity is computed with various distance metrics, where the sentences can be in the same or different languages. Spearman's correlation based on cosine similarity is the primary metric, evaluated against ground truth using Pearson and Spearman correlations, following the Sentence-BERT setting (Reimers and Gurevych 2019).

[IR] Retrieval. The task involves retrieving relevant documents from a corpus based on test queries. Each dataset includes a corpus, queries, and their relevant document mappings. Queries and documents are embedded using a model, and similarity scores are calculated with cosine similarity. The documents are ranked for each query, and evaluation metrics such as nDCG@k,

MRR@k, MAP@k, precision@k, and recall@k are computed, with nDCG@10 as the main metric.

[IR] Reranking. The input consists of a query and a set of both relevant and irrelevant reference texts. The objective is to rank these references based on their relevance to the query. The model generates embeddings for the reference texts, which are then compared to the query embedding using cosine similarity. Rankings are produced for each query and evaluated by averaging the results across all queries. Mean Average Precision (MAP) is used as the primary metric.

[TL] Multi-label Classification. This task entails predicting several labels for each input, where instances may belong to multiple classes. The model embeds the input data, and a multi-label classifier is applied on top of these embeddings. The F1 score is used as the main evaluation metric.

[IR] Instruction Retrieval. For instruction retrieval, the task expands upon traditional information retrieval by incorporating detailed instructions into queries. In contrast to standard retrieval, which typically uses short keyword queries, instruction retrieval pairs each query with a detailed instruction outlining the criteria for determining document relevance. Therefore, the task involves using each query's specific instruction, rather than a generic one, to retrieve relevant documents from the corpus. We follow Weller et al. (2024) using standard retrieval metrics with the instructions alongside the queries. nDCG@5 is used as the primary metric.

3.3 Ensuring Data Quality

Data quality assurance is achieved through systematic review processes by native speakers of SEA languages, all of whom are also proficient in English, who verified and validated the data for grammatical correctness, native written style, appropriate language usage (no code-switching), and correctness of the gold standard annotations. Only datasets that passed all criteria were included in our benchmark.

3.4 New Datasets

We also propose new datasets for Thai and Burmese on STS, NLI, and multi-label classification tasks. While Thai has multi-label classification datasets for evaluating embedding model robustness, no such datasets are available for Burmese. Importantly, there are also no available datasets for Thai and Burmese textual similarity and relation classification. As demonstrated in previous works (Gao, Yao, and Chen 2021; Chuang et al. 2022), these tasks also directly affect the performance of other tasks, such as retrieval and re-ranking. To address this problem, we propose 4 new datasets with 3,147 samples for Thai and 7 new datasets with 13,177 samples for Burmese.

As shown in Table 2, we used English sets of STS datasets (STSBenchmark (Cer et al. 2017), STS-2017 (Cer et al. 2017), STS-2022 (Chen et al. 2022), STS-2024 (Ousidhoum et al. 2024b), BIOSSES (Soğancıoğlu, "Ozt"urk, and "Ozg"ur 2017)), and NLI (XNLI (Conneau et al. 2018)) as the original texts. In addition, we translated a Thai multi-label classification dataset called Prachathai67k (cstorm125 2019) to Burmese. We use this setting because Thai and Burmese have similar cultures, politics, and histories; using Thai datasets as the starting dataset for Burmese translation is thus more suitable than using English datasets. Then, we asked Thai and Burmese native speakers (see Appendix 1.5 for annotator demographics) to translate the selected datasets in which the annotators were given the following instructions: *translate the selected datasets to make them a human-like or everyday conversation in your native languages* and *change the subject of a sentence to be gender-neutral* since both the Thai and Burmese languages have words or morphemes that can express the gender of the speaker. Therefore, the quality of our new human-crafted dataset is higher than that of using machine translations or LLMs to generate data, as such methods have been observed to be less native-like or unrepresentative of natural language

use (Lovenia et al. 2024; Singh et al. 2025). Note that we also evaluate the quality and differences between human-crafted and machine-translated datasets in Section 5.4.

Dataset	mya	tha
Biosses	100	100
STS17	250	250
STS22	197	197
STS24	2,600	2,600
STSBenchmark	2,880	-
XNLI	5,000	-
Prachathai67k	2,150	-
Total number of samples	13,177	3,147

Table 2: Statistics of the new evaluation datasets included in SEA-BED.

3.5 Benchmark Efficiency

Caching Embeddings. To improve the run-time efficiency, we use embedding caching to store embedded texts in memory and cache files; when seen texts are input to the same model, we will use the cached embedding instead of computing the new one to decrease the run-time of our benchmark.

Downsampling. Enevoldsen et al. (2025) proposed a downsampling technique for the English benchmark, decreasing the number of samples by 98%. However, as shown in Table 3, we applied the same technique to our benchmark (bitext mining datasets) and found that the performance of each model increased in all cases. This is because all challenging samples may have been removed from the dataset, leading to improved performance for most models. Moreover, the ranking of each model changed, in contrast to the findings of Enevoldsen et al. (2025), where the rankings remained largely unchanged. Therefore, we did not apply the downsampling technique to our benchmark.

Model	100% Dataset	30% Dataset	Rank after downsampling
multilingual-e5-large-instruct (560M)	87.86	93.03	0
Qwen3-Embedding-8B (8B)	84.78	90.31	↓1
bge-multilingual-gemma2 (9B)	82.02	90.71	↑3
multilingual-e5-large (560M)	84.51	88.19	↓1
bge-m3 (568M)	86.18	91.89	↑1
GritLM-7B (7B)	63.63	69.68	0
e5-mistral-7b-instruct (7B)	65.30	73.42	0
Qwen3-Embedding-0.6B (595M)	56.53	62.95	0
multilingual-mpnet-base (278M)	68.12	73.97	0
LaBSE (471M)	86.84	90.51	↓2
multilingual-MiniLM-L12 (118M)	53.23	59.06	0
Gemma-SEA-LION-v3-9B-IT (9B)	15.31	3.21	↓1
Sailor2-8B-Chat (8B)	4.31	6.01	↑1

Table 3: We evaluate 13 models on bitext mining using 100% and 30% dataset sizes. We also indicate the rank change of the model before and after downsampling to show the performance discrepancy.

3.6 Benchmark Discussion

Data Coverage. As shown in Table 1, our benchmark contains 169 datasets across 9 tasks for 10 SEA languages. As a point of comparison, we note that while MMTEB contains 270 datasets, only 21 are written in SEA languages, partly because there were few SEA researchers involved in its creation. In contrast, SEA-BED contains 169 datasets, where 148 datasets (∼87% of our benchmark) do not feature in MMTEB, making our benchmark more representative of SEA than MMTEB. In addition, 120 datasets in our benchmark were directly created by native speakers in the respective native languages rather than relying on machine translation or multilingual datasets. Moreover, we present 11 new datasets (STS, NLI, and multi-label classification tasks) for Thai and Burmese, enabling our benchmark to be evaluated for those languages and tasks.
Domain Coverage. We found that datasets in MMTEB only cover main domains (i.e., News, non-fiction, and encyclopedia) on SEA languages, while domains that are close to real-world use cases (i.e., social, legal, and medical) are not included in the benchmark. To address this problem, SEA-BED aims to represent more SEA and real-world use cases by covering 17 domains across 169 datasets, as shown in Table 4a. In addition, our benchmark also covers 4 new domains that never appeared in MMTEB (i.e., academic, blogs, medical, and subtitles). The full details of each domain can be found in Appendix 1.3
Task Coverage. As shown in Table 4b, MMTEB provides only limited task coverage for SEA languages, focusing on core tasks like cross-lingual pairing and topic clustering. In contrast, our benchmark proposes 9 new tasks, e.g., dialect pairing, written-forms pairing, language identification, toxic language detection, instruction QA, sentiment, topic classification, article reranking, and long document retrieval. This allows us to evaluate a broader range of tasks compared to previous benchmarks. Examples of each task can be found in Appendix 1.4.

Domain	ind	tha	vie	mya	fil	tam	khm	zsm	lao	tet
Academic			+							
Blog	+	+	+	+	+					
Constructed	✓	+	+	+	+			+		
Encyclopedia	✓	✓	✓	✓	✓	✓	✓	+	✓	
Fiction	+	✓	✓	+	✓	✓	+	+	+	
Government	+	✓	✓	+	+	✓	+	+	+	+
Legal	+	+	+	+		✓	+	+	+	
Medical		+	+							
News	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Non-fiction	✓	✓	✓	✓	✓	✓	✓	+	✓	
Religious	✓	✓		✓	✓	✓	+	+	+	+
Reviews	✓	✓	✓		+		+	+		
Social	+	+	+	+	✓	✓				
Spoken	✓	✓	✓	+	✓	✓	✓	✓	✓	+
Subtitles	+									
Web	✓	+	+	+		✓		+		+
Written	✓	✓	✓	✓	✓	✓	✓	+	✓	+

(a) Domain Coverage

Task	ind	tha	vie	mya	fil	tam	khm	zsm	lao	tet
Bitext Mining										
Cross-lingual pairing	✓	✓	✓	✓	✓	✓	✓	✓	✓	+
Dialect pairing	+	+	+	+	+		+	+	+	
Written-forms pairing	+	+	+	+						
Classification										
Language Identification	+	+	+	+	+		+	+	+	+
Sentiment	✓	✓	✓	+	✓	✓	+			
Topic Classification	✓	✓	✓	+	✓	✓	✓	✓	✓	
Toxic Language Detection	+	+	+		+					
Clustering										
Topic Clustering	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Instruction Retrieval										
Instruction QA	+	+	+							
Multi-label Classification										
Sentiment	+		+							
Topic Classification		+		+	+		+			
Toxic Language Detection	+									
Pair Classification										
Textual Entailment	✓	✓	✓	+	+	+	+	+	+	
Reranking										
Article Reranking	+	+								
Retrieval										
Article Retrieval	✓	✓	+	+						
Long Document Retrieval		+								
Question Answering	+	+	✓			+		+		
STS										
Multilingual STS	✓	+	+	+	+	+	+	+	+	
Cross-lingual STS		+		+		✓				

(b) Task Coverage

Table 4: Coverage of SEA-BED benchmark compared to MMTEB. For the above, + indicates newly added, while ✓is covered in MMTEB for SEA languages.

4. Experimental Settings

4.1 Models

To evaluate sentence embedding on SEA texts, we experiment on 13 open-source models across encoder and decoder models as follows:

- **multilingual-e5-large** (Wang et al. 2024b). A multilingual-e5-large model that is trained on over 100 languages using a combination of contrastive pre-training on diverse multilingual text pairs and supervised fine-tuning on high-quality labeled datasets using mined hard negatives and knowledge distillation techniques.
- **multilingual-e5-large-instruct** (Wang et al. 2024b). The multilingual-e5-large-instruct model is similar to multilingual-e5-large, with additional fine-tuning on instructional data.
- **e5-mistral-7b-instruct** (Wang et al. 2024a). The e5-mistral-7b-instruct model is a text embedding model based on Mistral-7B (Jiang et al. 2023), fine-tuned with contrastive learning on synthetic instruction data across 93 languages. Using a two-step prompting strategy, the model learns from diverse embedding tasks and achieves strong multilingual performance with under 1,000 training steps.
- **multilingual-mpnet-base-v2** (Reimers and Gurevych 2020). The multilingual-mpnet-base-v2 model is trained on parallel data for over 50 languages via multilingual knowledge distillation using paraphrase-mpnet-base-v2 (Reimers and Gurevych 2019) as a teacher model, xlm-roberta-base (Conneau et al. 2019) as a student model, and MSE loss to align their embeddings.
- **LaBSE** (Feng et al. 2022). The LaBSE model is trained on over 109 languages using a dual-encoder transformer architecture based on BERT (Devlin et al. 2018), leveraging a translation ranking loss function to produce sentence embeddings that align semantically similar sentences across languages into a shared vector space
- **multilingual-MiniLM-L12-v2** (Reimers and Gurevych 2020). The multilingual-MiniLM-L12-v2 model is trained using a similar multilingual knowledge distillation approach to multilingual-mpnet-base-v2, with paraphrase-MiniLM-L12-v2 (Reimers and Gurevych 2019) as a teacher model, Multilingual-MiniLM-L12-H384 (Wang et al. 2020) as a student model, and MSE loss to align their embeddings.
- **bge-m3** (Chen et al. 2024). The BGE-M3 model is trained on over 100 languages using a combination of contrastive pre-training on diverse multilingual corpora and supervised fine-tuning with high-quality labeled and synthetic datasets, leveraging hard negative mining and a self-knowledge distillation framework that integrates dense, sparse, and multi-vector retrieval signals.
- **bge-multilingual-gemma2** (Chen et al. 2024). The bge-multilingual-gemma2 model is built on Gemma-2-9b (Team 2024) and trained on diverse multilingual data across tasks such as retrieval, classification, and clustering using embedding techniques.
- **GritLM-7B** (Muennighoff et al. 2024b). The GritLM-7B model is built on the Mistral-7B (Jiang et al. 2023) architecture and trained using Generative Representational Instruction Tuning (GRIT), a unified framework combining contrastive learning for embeddings and next-token prediction for generation, with task-specific instructions and a joint loss to enable strong performance across both tasks.
- **Qwen3-Embedding-0.6B and Qwen3-Embedding-8B** (Zhang et al. 2025). The Qwen3-Embedding-0.6B and Qwen3-Embedding-8B models were trained on multiple languages using a multi-stage training pipeline that combines large-scale weakly supervised pre-training on synthetic multilingual data with supervised fine-tuning and model merging techniques to enhance robustness and generalization.

- **Sailor2-8B-Chat** (Dou et al. 2025). The Sailor2-8B-Chat model, based on an expanded Qwen2.5-7B (Yang et al. 2024), was trained on 13 SEA languages using two-stage continual pre-training with balanced and high-quality data, followed by two-stage instruction tuning and preference tuning with length-regularized DPO (Park et al. 2024).
- **Gemma-SEA-LION-v3-9B-IT** (Singapore 2024). The Gemma-SEA-LION-v3-9B-IT model is fine-tuned from the Gemma2 9B (Rivière et al. 2024) base model on English and multiple SEA languages (such as Indonesian, Thai, and Vietnamese), using a combination of full parameter fine-tuning, on-policy alignment, and model merging techniques.

Moreover, we also evaluate the performance of proprietary models as follows:

- **text-embedding-3-small**. We evaluate the text-embedding-3-small [3] model, which provides a highly efficient embedding model suitable for various downstream applications.
- **embed-multilingual-v3.0**. We evaluate the embed-multilingual-v3.0 [4] model, designed for multilingual representation learning across over 100 languages.
- **voyage-3**. We evaluate the voyage-3 [5] model, which provides efficient, high-quality embeddings optimized for retrieval across diverse domains.
- **jina-embeddings-v3**. We evaluate the jina-embeddings-v3 (Sturua et al. 2024b) model, which is designed for efficient semantic similarity and search applications, supporting various multilingual scenarios.

All proprietary models were accessed and evaluated using their latest publicly available versions during experimentation (April 4th, 2025).

4.2 Evaluation Setup

We utilize the evaluation metrics of each task as mentioned in Section 3.2. We use the averaging strategy similar to previous works (Muennighoff et al. 2023; Enevoldsen et al. 2025), averaging all the tasks equally with the standard deviation (SD) score. We acknowledge that the metrics for each task are different (e.g., F1 for classification and nDCG@10 for retrieval). Thus, we provide the analysis for both individual and average results instead of focusing only on the average score. All experiments were run on eight H100 (80 GB).

5. Experimental Results

In this section, we present a series of studies using our SEA sentence embedding benchmark. We evaluate embedding models across tasks and languages in Section 5.1 and Section 5.2, respectively. Section 5.3 compares the changes in rankings and embedding models' performances when evaluated on SEA-BED versus when evaluated on other multilingual sentence embedding benchmarks. Section 5.4 studies the correctness and effectiveness of machine translation and human translation datasets. We also study the correlation between the tokenizer and model performance in Section 5.5. Moreover, we analyze language similarities in SEA, as detailed in Section 5.6.

5.1 Main Results

We begin with an overall task-based performance analysis (RQ1), asking which tasks remain particularly challenging for state-of-the-art models across SEA languages. We evaluate each

3 https://openai.com/index/new-embedding-models-and-api-updates
4 https://cohere.com/blog/introducing-embed-v3
5 https://blog.voyageai.com/2024/09/18/voyage-3/

model on the seven tasks in SEA-BED and compare average scores. In particular, we evaluate 13 open-source models and 4 proprietary models on 169 datasets, where each task's evaluation metrics, task details, and desired properties are discussed in Section 3.2.

Results. As shown in Table 5, the experiment results demonstrate that multilingual-e5-large-instruct performs the best on our benchmark, achieving 75.24 points on the average score. The performance of the second-best model (Qwen3-Embedding-8B) is lower than that of multilingual-e5-large-instruct by only 0.06 points on average, with a 70 times difference in the model parameters (560M vs. 8B parameters). Moreover, we found that, although Gemma-SEA-LION-v3 and Sailor2 were trained for SEA languages specifically, the models did not perform well on our sentence embedding benchmark. However, this is perhaps unsurprising since these models were never trained for sentence embedding purposes. For the proprietary models, in contrast to the English sentence embedding benchmark (Muennighoff et al. 2023) findings that proprietary models outperformed open-source models, we found that all proprietary models perform lower than multilingual-e5-large-instruct and Qwen3-Embedding-8B. This indicates that all proprietary models might be trained primarily in English and not optimized for SEA languages.

Model	**Dim.**	**Btxt**	**Clf**	**Clust**	**In. Rtrvl**	**M. Clf**	**Pr. Clf**	**Rtrvl**	**Rrnk**	**STS**	**Avg.**
		SEA-BED									
Number of datasets (→)		(26)	(73)	(10)	(4)	(11)	(13)	(20)	(1)	(11)	(169)
Open-source											
multilingual-e5-large-instruct (560M)	1024	**87.86**	77.70	**58.09**	69.10	87.84	66.58	77.16	77.24	**75.59**	**75.24**$_{\pm 9.06}$
Qwen3-Embedding-8B (8B)	4096	84.78	78.60	52.93	70.81	90.57	63.10	**81.99**	78.51	75.31	75.18$_{\pm 10.84}$
bge-multilingual-gemma2 (9B)	3584	82.02	78.13	49.14	**71.52**	**90.89**	**73.87**	80.55	69.04	72.53	74.19$_{\pm 10.85}$
multilingual-e5-large (560M)	1024	84.51	**78.24**	47.83	66.06	88.94	65.79	78.25	**79.00**	69.61	73.14$_{\pm 11.66}$
bge-m3 (568M)	4096	86.18	75.98	42.23	58.51	89.89	68.73	73.56	75.98	73.27	71.59$_{\pm 13.48}$
GritLM-7B (7B)	4096	63.63	77.47	46.29	67.60	88.76	63.86	65.97	73.37	64.69	67.96$_{\pm 10.92}$
e5-mistral-7b-instruct (7B)	4096	65.30	76.65	49.48	54.46	88.32	63.81	72.93	75.33	63.50	67.75$_{\pm 11.24}$
Qwen3-Embedding-0.6B (595M)	1024	56.53	74.47	43.94	65.80	88.19	60.36	76.24	75.03	65.74	67.37$_{\pm 11.58}$
multilingual-mpnet-base (278M)	768	68.12	73.79	41.12	52.44	87.28	70.79	58.28	64.01	70.15	65.11$_{\pm 12.55}$
LaBSE (471M)	768	86.84	75.19	41.39	39.73	86.65	62.32	53.72	61.23	68.32	63.93$_{\pm 16.31}$
multilingual-MiniLM-L12 (118M)	768	53.23	70.50	31.50	48.66	84.88	65.70	52.47	62.27	64.59	59.31$_{\pm 14.25}$
Gemma-SEA-LION-v3-9B-IT (9B)	3584	15.31	75.87	39.94	11.02	89.94	57.77	22.03	65.49	38.85	46.25$_{\pm 26.18}$
Sailor2-8B-Chat (8B)	3584	4.31	76.43	38.51	3.29	90.21	56.71	10.09	47.05	37.25	40.43$_{\pm 29.25}$
Proprietary models											
embed-multilingual-v3.0	1024	**88.32**	**78.52**	48.99	65.59	**89.98**	**66.11**	**78.17**	**77.77**	73.11	**74.06**$_{\pm 11.89}$
jina-embeddings-v3	1024	81.86	77.40	**50.90**	**69.11**	88.97	63.61	76.28	72.49	**73.17**	72.64$_{\pm 10.30}$
voyage-3	1024	55.62	75.72	45.15	61.77	88.70	60.23	62.91	74.62	61.97	65.19$_{\pm 12.01}$
text-embedding-3-small	1536	43.12	72.88	39.34	52.87	88.19	60.16	65.18	71.25	52.31	60.59$_{\pm 14.65}$

Table 5: SEA-BED: Evaluation results across different tasks.

Discussion. We found that task performance consistency is the main challenge for current sentence embedding models. In particular, a robust model should perform well on all tasks. As shown in Table 5, we found that there is no dominant model that achieves the highest score on all tasks. Notably, model performance varies considerably depending on the task. For example, the proprietary model embed-multilingual-v3.0 excels in Bitext Mining (Btxt) and Classification (Clf) tasks, achieving the top scores among all evaluated models in these categories. In contrast, open-source models such as bge-multilingual-gemma2 achieve the highest score on Instruction Retrieval (In. Rtrvl), Multilingual Classification (M. Clf), and Pair Classification (Pr. Clf) tasks, while multilingual-e5-large-instruct performs well on Clustering (Clust) and STS tasks. Additionally, Qwen3-Embedding-8B shows strong performance on the Retrieval (Rtrvl) task. Moreover, the second-best performer models (highlighted as underscore) are various and diverse in model size, ranging from 278 million parameters to 9 billion parameters. *This emphasizes that the task consistency problem in our benchmark is still challenging for embedding models*. In conclusion,

when using multilingual sentence embedding in SEA languages, we need to select the model based on the task at hand, and there is no all-rounder model for every solution.

5.2 Language Breakdown

This study presents a language-wise analysis for RQ1 to pinpoint which SEA languages and scripts see the largest performance drops. We use the same metric and datasets; however, the number of datasets will be higher than in the previous study because one dataset can contain more than one language. Therefore, in total, there are 294 datasets for this study.

Results. As shown in Table 6, we observe performance variation across languages that, when compared to the task-level results in Table 5, suggests that strong overall performance does not necessarily imply consistent multilingual coverage. For example, while multilingual-e5-large-instruct achieves the highest average score overall (78.93), its performance varies across languages, ranging from 69.40 points in Tetum to 84.60 points in Malay. Similarly, we also observe inconsistent performance in Qwen3-Embedding-8B, which ranks as the second-best model overall, demonstrates strong performance in high-resource SEA languages such as Thai (81.49 points) and Vietnamese (78.99 points), but underperforms in lower-resource SEA languages like Tetum (67.44 points). Moreover, we also found that *some models do not fully support SEA languages*, e.g., GritLM-7B does not support Burmese, Khmer, and Lao, while bge-multilingual-gemma2 does not support Lao. Although those models were found to perform well in this experimental study, the fact that they do not support some SEA languages results in those models being less appropriate for real-world applications for SEA languages.

Model	**ind**	**tha**	**vie**	**mya**	**fil**	**tam**	**khm**	**zsm**	**lao**	**tet**	**Avg.**
					SEA-BED						
Number of datasets (→)	(70)	(55)	(40)	(33)	(28)	(18)	(18)	(14)	(14)	(4)	(294)
Open-source											
multilingual-e5-large-instruct (560M)	79.50	81.11	78.00	**78.37**	79.19	77.09	**78.13**	**84.60**	**83.94**	69.40	**78.93**$_{\pm 3.98}$
Qwen3-Embedding-8B (8B)	79.73	**81.49**	**78.99**	74.91	78.05	75.95	75.46	82.39	78.20	67.44	77.26$_{\pm 4.02}$
bge-multilingual-gemma2 (9B)	79.93	80.58	78.76	70.01	**79.61**	**80.96**	74.39	83.38	65.82	65.05	75.85$_{\pm 6.31}$
multilingual-e5-large (560M)	78.59	79.89	78.93	70.28	77.98	77.83	72.11	80.10	79.91	63.55	75.92$_{\pm 5.22}$
bge-m3 (568M)	78.09	77.59	75.91	73.12	75.78	77.51	76.23	82.54	82.26	65.53	76.46$_{\pm 4.55}$
GritLM-7B (7B)	**80.47**	72.84	77.37	45.05	77.49	60.42	52.58	78.41	30.07	**69.67**	64.44$_{\pm 16.13}$
e5-mistral-7b-instruct (7B)	79.23	74.77	75.37	48.85	78.10	66.73	56.49	78.82	27.99	66.73	65.32$_{\pm 15.74}$
Qwen3-Embedding-0.6B (595M)	75.60	75.85	75.13	49.08	63.11	61.12	44.10	69.51	29.78	63.38	60.67$_{\pm 14.55}$
multilingual-mpnet-base (278M)	74.60	73.91	72.66	61.19	52.02	63.31	64.44	75.48	65.63	50.78	65.40$_{\pm 8.53}$
LaBSE (471M)	73.98	70.20	72.60	73.63	76.99	76.59	74.06	82.87	79.84	69.11	74.99$_{\pm 3.99}$
multilingual-MiniLM-L12 (118M)	71.48	70.42	69.90	54.48	47.28	27.88	39.92	69.58	45.34	47.69	54.40$_{\pm 14.53}$
Gemma-SEA-LION-v3-9B-IT (9B)	49.86	41.67	51.90	30.80	54.14	29.20	39.53	49.24	22.01	25.06	39.34$_{\pm 11.27}$
Sailor2-8B-Chat (8B)	49.54	35.98	42.94	30.14	46.16	28.57	28.57	30.75	18.31	25.76	33.67$_{\pm 9.33}$
Proprietary models											
embed-multilingual-v3.0	**79.72**	**80.99**	**78.93**	**76.13**	**78.99**	**78.87**	**77.01**	**82.42**	**83.34**	**66.76**	**78.32**$_{\pm 4.39}$
jina-embeddings-v3	77.35	78.64	76.10	75.10	74.25	76.14	74.73	77.91	77.91	65.11	75.32$_{\pm 3.68}$
voyage-3	75.56	69.78	73.68	48.19	71.43	67.28	35.02	69.13	24.27	61.48	59.58$_{\pm 16.83}$
text-embedding-3-small	78.34	55.24	70.06	32.79	68.08	35.38	30.15	69.78	23.97	65.09	52.89$_{\pm 19.18}$

Table 6: SEA-BED: Evaluation results across each language.

Discussion. Experimental results demonstrate the language consistency problem, where open-source models perform inconsistently on each language. Although multilingual-e5-large-instruct might perform best in the overall performance, we found that no model can perform best for all languages. multilingual-e5-large-instruc performs well on Burmese, Khmer, Malay, and Lao, while Qwen3-Embedding-8B performs well on Thai and Vietnamese, bge-multilingual-

gemma2 performs well on Filipino and Tamil, and GritLM-7B performs well on Indonesian and Tetum. This emphasizes the "Language Consistency" problem, where all languages perform inconsistently, unsteadily, and inconclusively. In contrast, we found that embed-multilingual-v3.0 consistency outperforms all proprietary models in all languages. Although the overall performance of proprietary models is lower than open-source models, when it comes to real-world applications that support multiple SEA languages, using proprietary models might be more reliable than open-source models.

5.3 Performance Changes Analysis

Here, we address RQ2, examining how SEA-focused performance contrasts with the broader multilingual benchmark (MMTEB). To study the robustness of embeddings in world and SEA languages, we compare the ranking changes between our benchmark and the recent multilingual sentence embedding benchmark, MMTEB. We use the task average metric (Table 5), similar to MMTEB.

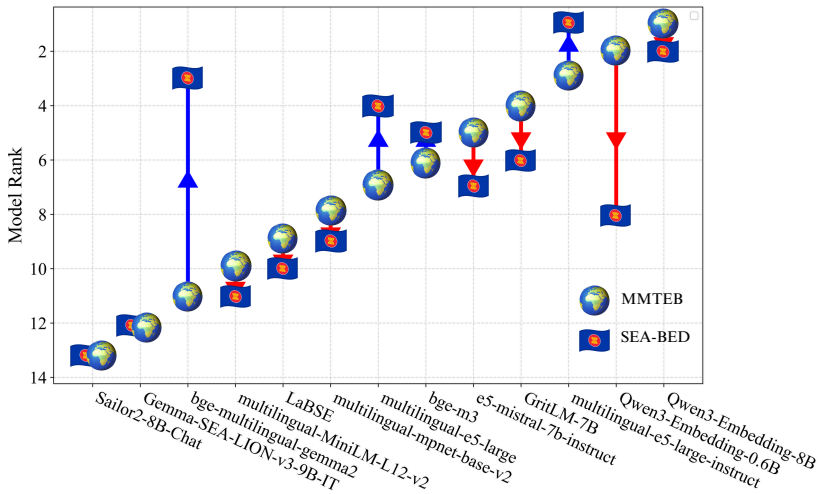


Figure 2: Ranking difference between MMTEB and SEA-BED.

As shown in Figure 2, based on the experiment from MMTEB, Qwen3-Embedding-8B performed the best on world results, which includes 1,090 languages[6]. However, when we focus only on SEA languages using SEA-BED, the ranking of Qwen3-Embedding-8B dropped from the first rank to the second rank. In addition, Qwen3-Embedding-0.6B dropped from second rank to eighth rank. This is because some of the linguistic and dialect knowledge will be different compared to other groups of languages, when we evaluate them only for the SEA languages. Moreover, the proportion of training data is also a factor since the portion of SEA training data in Qwen3-Embedding might be smaller compared to other languages. On the other hand, the rankings of multilingual-e5-large-instruct, bge-multilingual-gemma2, and multilingual-e5-large increased significantly. This emphasizes that the challenge, gaps, and model capabilities measured in MMTEB and our benchmark differ, especially in the supported languages of embedding models that do not fully support SEA languages. Even though some models are used to perform well on MMTEB, they are not guaranteed to achieve the same performance for SEA languages. We

6 We obtained the model rankings on June 11th, 2025.

therefore appeal to the NLP community to develop embedding models to support more SEA languages.

5.4 Machine-Crafted vs. Human-Crafted Datasets

This subsection addresses RQ3 by testing whether human-crafted data yields results different from machine-generated data. We split the experiment into machine generation and translation studies.

Machine Translation vs. Human-crafted Datasets To observe the differences between machine translation and human datasets, we conduct a study using our newly proposed datasets for Thai and Burmese STS tasks. In the process of creating SEA-BED, we have asked all of our annotators to translate from English to Thai and Burmese without relying on machine translations, hence enabling us to study the differences between MT and human-created datasets. For MT datasets, we use the same example originally written in English and translate it into Burmese and Thai using Google NMT (accessed February 15th 2025).

Interestingly, as shown in Table 7, the difference in performance on Thai datasets is small for both MT and human-annotated sets. We can observe that the performance difference is lower than 2 Spearman's correlation points for all cases. This corroborates findings from previous English-Thai machine translation works, which found that the current Google NMT for English-Thai is good enough for use in real-world scenarios (Lowphansirikul et al. 2022; Chiaranaipanich et al. 2024). However, satisfactory performance may be limited by the domain. That is, machine translation is observed to yield robust results since our data is within the general domain. In contrast, translating texts written in the medical or legal domain might yield different results, as demonstrated in previous Thai machine translation works (Pengpun et al. 2024a; Chiaranaipanich et al. 2024).

In contrast to Thai, the performance gap between Burmese human and machine translation datasets is larger than that of Thai in most cases. We found that the Google NMT results for Burmese sometimes show code-switching between Thai and Burmese characters. As an example, Figure 3 shows that the Google NMT output is Burmese that has Thai mixed in. This emphasizes that, in underrepresented languages, using humans to create evaluation datasets is still better than relying on machine translations. Using machine translations might be appropriate for high-resource languages in SEA (i.e., Thai) since it can produce comparable results to the human-crafted dataset, but when the languages are underrepresented (i.e., Burmese), using humans is still empirically better than machine translations.

Model	Original (eng)	Machine (mya)	Human (mya)	Machine (tha)	Human (tha)
multilingual-e5-large-instruct (560M)	82.87	74.82	75.06	79.66	79.80
Qwen3-Embedding-8B (595M)	81.17	74.02	75.81	80.75	80.74
bge-multilingual-gemma2 (9B)	84.64	75.51	72.87	80.25	78.97
multilingual-e5-large (560M)	80.00	71.49	71.55	76.44	76.50
bge-m3 (568M)	80.86	74.57	71.96	77.75	76.07
GritLM-7B (7B)	82.65	65.60	66.03	74.64	74.81
e5-mistral-7b-instruct (7B)	81.86	62.36	64.63	74.57	74.57
Qwen3-Embedding-0.6B (8B)	80.11	67.10	69.23	77.88	77.79
multilingual-mpnet-base (278M)	80.54	72.34	71.16	72.61	72.60
LaBSE (471M)	73.50	69.06	70.04	69.29	68.83
multilingual-MiniLM-L12 (118M)	78.89	69.27	67.26	72.25	72.23
Gemma-SEA-LION-v3-9B-IT (9B)	60.42	46.50	49.29	55.97	56.01
Sailor2-8B-Chat (8B)	57.94	48.79	52.89	55.85	54.36

Table 7: Model performance on Machine Translation vs. Human Datasets on our STS datasets (Table 2).

เจริญปราดซ์ลม์းမုတစ်ဆင့် သမ္မတရုံးခန်းမသို့ ဝင်ရောက်သည်။
(Pass Charoen Pradit Road, enter the Office of the President's Auditorium.)

Periodic Table ၏ ဘယ်ဘက်ရှိ ဓာတုဒြပ်စင်များသည် အိုင်ယွန်ငါးရှင်း စွမ်းအင်ထက် များစွာနိမ့်ကျသည်။
(The chemical elements to the left of the periodic table have a much lower ionization energy.)

Figure 3: Code-switching between Thai and Burmese words (translated by Google NMT).

Machine Generation vs. Human Datasets Nowadays, many research papers propose a dataset and benchmark that rely on machine learning, due to the scalability of automation. However, we argue that the quality of such purely machine-generated data is unstable and should not be the majority dataset in the benchmark, since it will affect the performance and results of studies. To confirm this, we conduct a study by comparing human-crafted and non-human-crafted datasets using the top five models from our previous studies. In particular, we perform the same task and language but select only datasets that are formulated by humans or machines. We demonstrate the performance of embedding models on human and non-human data in Tables 8 and 9, respectively.

Model	ind	tha	vie	mya	fil	tam	khm	zsm	lao	tet	Avg.
multilingual-e5-large-instruct (560M)	82.06	**84.00**	79.54	**78.27**	80.33	76.89	**79.63**	88.68	**87.07**	**41.06**	**77.33**$_{\pm 12.50}$
Qwen3-Embedding-8B (8B)	82.98	83.61	**80.67**	75.35	78.78	75.66	75.86	86.87	80.70	36.00	75.65$_{\pm 13.70}$
bge-multilingual-gemma2 (9B)	**83.28**	82.90	80.29	70.43	**80.67**	**80.99**	74.03	85.18	66.50	30.91	73.52$_{\pm 15.32}$
multilingual-e5-large (560M)	81.90	82.54	**80.67**	70.56	79.42	78.03	72.63	82.92	82.81	32.24	74.37$_{\pm 14.63}$
bge-m3 (568M)	81.28	80.45	77.48	73.70	76.87	77.70	76.56	**89.21**	85.03	36.91	75.52$_{\pm 13.57}$

Table 8: The top five models evaluation results across each language on human-crafted datasets only.

Model	ind	tha	vie	mya	fil	tam	khm	zsm	lao	tet	Avg.
multilingual-e5-large-instruct (560M)	**72.04**	64.14	66.92	**79.47**	68.54	80.45	71.38	69.31	67.23	97.73	**73.72**$_{\pm 9.43}$
Qwen3-Embedding-8B (8B)	69.95	**68.99**	66.91	70.19	**71.24**	**80.86**	73.66	65.62	64.85	98.89	73.12$_{\pm 9.63}$
bge-multilingual-gemma2 (9B)	71.17	66.97	**67.80**	65.55	69.77	80.40	**76.00**	**76.63**	62.19	**99.18**	73.57$_{\pm 10.04}$
multilingual-e5-large (560M)	68.58	64.35	66.43	67.30	64.61	74.34	69.77	69.52	64.45	94.86	70.42$_{\pm 8.66}$
bge-m3 (568M)	69.14	60.77	64.57	66.90	65.62	74.39	74.74	61.77	**67.47**	94.15	69.95$_{\pm 9.18}$

Table 9: The top five models evaluation results across each language on machine-generated datasets only.

As can be seen, we observe two major changes in terms of performance: (i) average performance changes and (ii) model ranking changes. For the first problem, we observe that the overall performance of machine-generated datasets is always lower than that of human-crafted datasets, except for bge-multilingual-gemma2. This is also consistent with our previous study in machine translation datasets, as illustrated in Table 7, which shows that using machines to generate data will decrease the model's performance. For the second problem, the most problematic in the evaluation research work, the results are inconclusive because the model rankings change. A robust benchmark should produce the most correct results that align with human results, where using only machine-generated datasets might not address this desired property. The results from Tetum in Table 9 are also inconclusive since the performance of Tetum increased from 30.91 to 99.18 points using bge-multilingual-gemma2. This is because the dataset of machine-generated in

Tetum is an easy task, the language detection task using the MADLAD-400 dataset, where all models achieve more than 90 points. We assume that this is because the machine-generated data might be leaked to those models, or the data might be in-domain data for Tetum (those models have seen this test data or similar data before).

5.5 Tokenizer Analysis

This section revisits RQ1, exploring whether limited SEA-vocabulary coverage in multilingual tokenizers correlates with poor downstream results. We highlight scripts like Lao or Khmer, which are often underrepresented in tokenizers. Previous works (Ali et al. 2024; Arnett and Bergen 2025; Liang et al. 2023) demonstrated that vocabularies in a tokenizer affect the model performance in downstream tasks. In particular, when the multilingual tokenizer represents more vocabulary in some languages, the performance on those languages has also been observed to improve. In this study, we want to investigate whether the vocabulary in the tokenizer affects SEA-BED's overall performance or not. To answer this question, we count the SEA tokens in each sentence embedding model and compare their performance from Table 6.

As shown in Table 10, the language with the most tokens represented in a tokenizer is Filipino, with an average of 2.94 percent of vocabulary tokens in 13 models. However, compared to the language performance (Table 6), Filipino performance is lower than Indonesian. Surprisingly, there are no tokens for Tetum at all in the 13 models. We observe that performance on Tetum is also the worst compared to other SEA languages. Moreover, the performance is mixed for languages that do not use Latin characters, i.e., Thai, Burmese, Lao, and Tamil.

Model	ind	tha	vie	mya	fil	tam	khm	zsm	lao	tet
multilingual-e5-large-instruct (560M)	1.20	1.61	0.73	0.91	3.59	0.98	0.66	0.20	0.56	0.00
Qwen3-Embedding-8B (8B)	0.39	1.70	0.84	0.02	1.13	0.02	0.03	0.11	0.02	0.00
bge-multilingual-gemma2 (9B)	0.59	0.50	0.55	0.45	3.04	0.13	0.03	0.11	0.02	0.00
multilingual-e5-large (560M)	1.20	1.61	0.73	0.91	3.59	0.98	0.66	0.20	0.56	0.00
bge-m3 (568M)	1.20	1.61	0.73	0.91	3.59	0.98	0.66	0.20	0.56	0.00
GritLM-7B (7B)	0.27	0.19	0.55	0.45	3.04	0.13	0.03	0.11	0.02	0.00
multilingual-mpnet-base (278M)	1.20	1.61	0.73	0.91	3.59	0.98	0.66	0.20	0.56	0.00
LaBSE (471M)	1.12	0.45	0.81	0.45	4.65	1.28	0.54	0.19	0.29	0.00
e5-mistral-7b-instruct (7B)	0.27	0.19	0.55	0.45	3.04	0.13	0.03	0.11	0.02	0.00
Qwen3-Embedding-0.6B (595M)	0.39	1.70	0.84	0.02	1.13	0.02	0.03	0.11	0.02	0.00
multilingual-MiniLM-L12 (118M)	1.20	1.61	0.73	0.91	3.59	0.98	0.66	0.20	0.56	0.00
Gemma-SEA-LION-v3-9B-IT (9B)	0.59	0.50	0.55	0.45	3.04	0.13	0.03	0.11	0.02	0.00
Sailor2-8B-Chat (8B)	0.39	1.70	0.84	0.02	1.13	0.02	0.03	0.11	0.02	0.00
Average	0.77	1.15	0.71	0.53	2.94	0.52	0.31	0.15	0.25	0.00

Table 10: The percentage number of vocabulary tokens for each model in each language.

Performance Analysis. Additionally, we analyze the correlation between the percentage of vocabulary token coverage and performance scores for the two top-performing and two lowest-performing embedding models, as shown in Figure 4. The results indicate that the vocabulary size of each model does not have a direct effect on model performance in the sentence embedding benchmark for SEA languages. Although some models have a larger number of tokens in their tokenizers covering SEA vocabularies, their performance in the benchmark is not significantly higher than that of models with lower vocabulary coverage. This indicates that simply increasing vocabulary size does not necessarily lead to better performance in sentence embedding tasks for SEA languages.

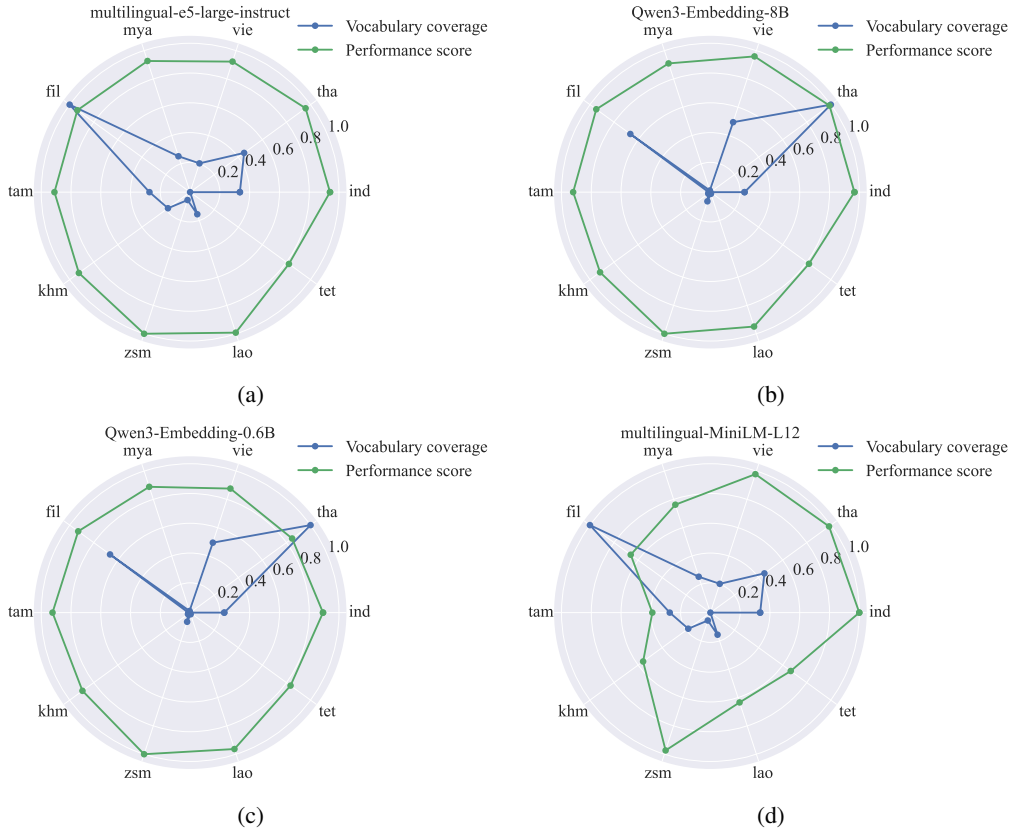


Figure 4: (Top) Correlation between the percentage of vocabulary token coverage and performance score for the two top-performing models, multilingual-e5-large-instruct and Qwen3-Embedding-8B. (Bottom) Correlation between the percentage of vocabulary token coverage and performance score for the two lowest-performing models, Qwen3-Embedding-0.6B and multilingual-MiniLM-L12. Both values are normalized to a [0, 1] scale for comparability across languages and models.

Discussion. In contrast to previous works, we summarize that the number of tokens present in the tokenizer might not strongly correlate with the performance in a language. There are many SEA languages with diverse scripts, and solely having a larger vocabulary for each language might not necessarily yield significant improvement. As shown in the language performances of GritLM-7B and bge-multilingual-gemma2 (Table 6), omitting SEA languages from the training data results in poor performance in those languages. To achieve a promising result, we can add more SEA training datasets in the training step to improve downstream task performance rather than adding more tokens in the tokenizer.

5.6 Language Similarity

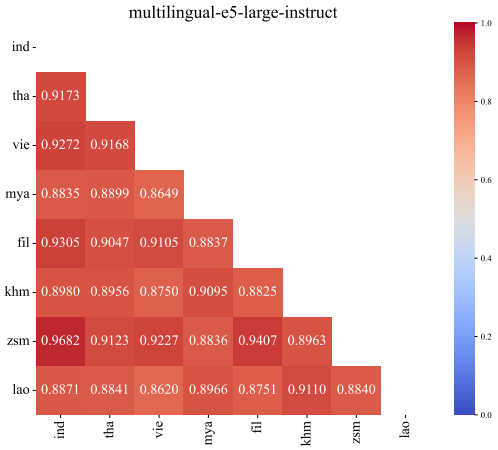
To further understand the similarity between language and performance (**RQ1**), we analyze the performance of bi-text retrieval datasets in SEA languages. In particular, we study the language similarity of robust and non-robust models, e.g., top-performing and worst-performing embedding models, to see what the desired property is to improve our benchmark. We utilize the dialect pairing subset task in this experiment, where we use a batch size of 128 for the negative pair

evaluation. In addition, we use cosine similarity as the main metric, where higher values indicate greater embedding similarity between language pairs.

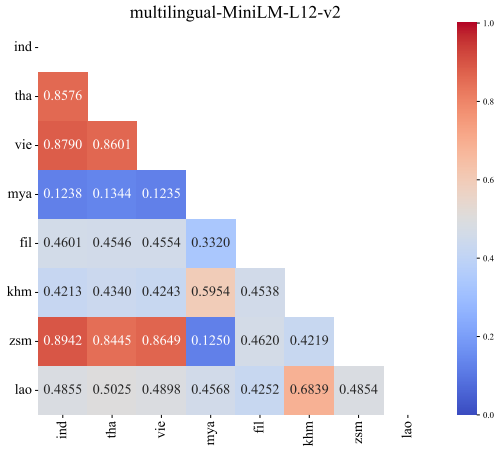
As shown in Figure 5, the top-performing model, multilingual-e5-large-instruct, shows consistently high similarity for positive samples, especially Indonesian-Malay (0.9682 points), Indonesian-Filipino (0.9305 points), and Thai-Vietnamese (0.9168 points), indicating strong cross-lingual embeddings. However, multilingual-e5-large-instruct unexpectedly maintains high similarity for negative samples (0.75–0.81 points), indicating limited distinction between unrelated sentence pairs and highlighting a gap for improvement. In contrast, multilingual-MiniLM-L12-v2 struggles with related positive pairs, showing lower similarity for Indonesian-Filipino (0.4601 points) and notably weak similarity with Burmese (around 0.12–0.59 points). Interestingly, this model achieves low similarity for negative pairs, mostly under 0.08 points, clearly distinguishing unrelated samples. Although it falls short in overall embedding quality, multilingual-MiniLM-L12-v2's distinct negative sample separation provides valuable insights into desirable characteristics for embedding models. These findings suggest that a balanced approach, achieving both strong cross-lingual similarity for positive examples and clear differentiation for negative examples, is essential to improve future embedding benchmarks.

6. Conclusion

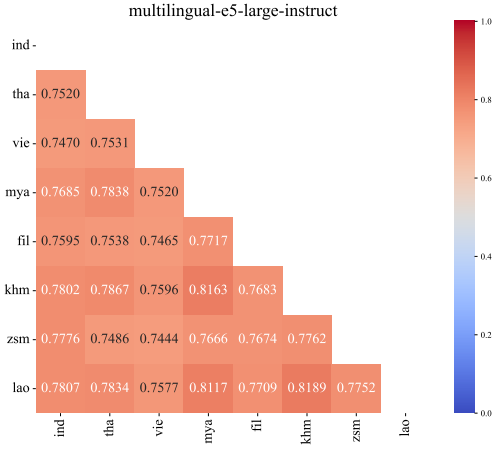
In this work, we present the Southeast Asian Massive Sentence Embedding Benchmark (SEA-BED). We experimented on 17 multilingual embedding models and 6 studies to reveal the challenges of our benchmark compared to previous sentence embedding benchmarks. The experiment of our studies reveals the challenge in SEA embeddings as follows: **Performance across different tasks** (Section 5.1): A robust model should perform well regardless of task, whereas current models favor only some tasks. **Performance across each language** (Sections 5.2 and 5.5): We found that some models do not fully support SEA languages, and the performance on each language is inconsistent. **Robustness in SEA and non-SEA languages** (Section 5.3): A model should perform well regardless of languages or benchmarks, but current embedding models cannot achieve high scores for both the world and SEA embedding benchmarks. Section 5.4 shows the possibilities and gaps of using machine translations and generations to formulate datasets in SEA languages. We also studied the correlation between vocabulary and downstream task performances to reveal the improvement manner for future work in Section 5.5. Lastly, we conducted a study to understand the language similarities in SEA for future embedding works to understand how the nuances present in the linguistic properties of SEA language datasets affect model performance on SEA-BED in Section 5.6.



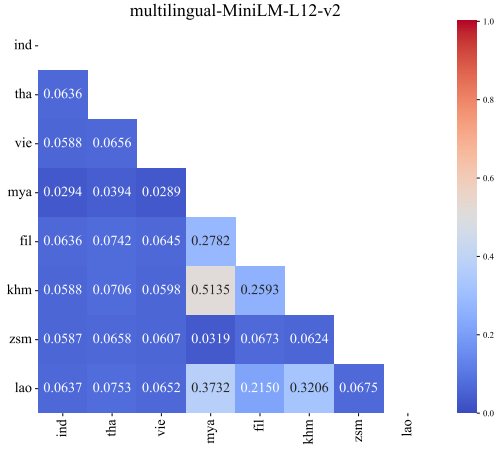
(a) The top-performing model on the positive pairs



(b) The worst-performing model on the positive pairs



(c) The top-performing model on the negative pairs



(d) The worst-performing model on the negative pairs

Figure 5: We perform cross-lingual similarity using the bitext mining task (dialect pairing subset). (Top) Cross-lingual similarity metrics of the top-performing and worst-performing embedding models on the positive parallel samples. (Bottom) Cross-lingual correlation metrics of the top-performing and worst-performing embedding models on the negative parallel samples.

References

A. N. Azhar, M. L. Khodra and A. P. Sutiono. Multi-label aspect categorization with convolutional neural networks and extreme gradient boosting.

Adelani, David Ifeoluwa, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and Annie En-Shiun Lee. 2023. Sib-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects.

Adelani, David Ifeoluwa, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and Annie En-Shiun Lee. 2024. Sib-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects.

Aggarwal, Divyanshu, Vivek Gupta, and Anoop Kunchukuttan. 2022. Indicxnli: Evaluating multilingual inference for indian languages.

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

Akarajaradwong, Pawitsapak, Pirat Pothavorn, Chompakorn Chaksangchaichot, Panuthep Tasawong, Thitiwat Nopparatbundit, and Sarana Nutanong. 2025. Nitibench: A comprehensive studies of llm frameworks capabilities for thai legal question answering.

Akerman, Vesa, David Baines, Damien Daspit, Ulf Hermjakob, Taeho Jang, Colin Leong, Michael Martin, Joel Mathew, Jonathan Robie, and Marcus Schwarting. 2023. The ebible corpus: Data and model benchmarks for bible translation for low-resource languages. *arXiv preprint arXiv:2304.09919*.

Alfina, Ika, Rio Mulia, Mohamad Ivan Fanany, and Yudo Ekanata. 2017. Hate speech detection in the indonesian language: A dataset and preliminary study.

Ali, Mehdi, Michael Fromm, Klaudia Thellmann, Richard Rutmann, Max Lübbering, Johannes Leveling, Katrin Klug, Jan Ebert, Niclas Doll, Jasper Schulze Buschhoff, Charvi Jain, Alexander Arno Weber, Lena Jurkschat, Hammam Abdelwahab, Chelsea John, Pedro Ortiz Suarez, Malte Ostendorff, Samuel Weinbach, Rafet Sifa, Stefan Kesselheim, and Nicolas Flores-Herr. 2024. Tokenizer choice for LLM training: Negligible or crucial? In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 3907–3924, Association for Computational Linguistics.

Arfinda Ilmania, Samuel Cahyawijaya Ayu Purwarianti, Abdurrahman. 2018. Aspect detection and sentiment classification using deep neural network for indonesian aspect-based sentiment analysis. In *Proceedings of the 2018 International Conference on Asian Language Processing(IALP)*, pages 62–67, IEEE.

Arnett, Catherine and Benjamin Bergen. 2025. Why do language models perform worse for morphologically complex languages? In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 6607–6623, Association for Computational Linguistics.

Artetxe, Mikel, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856.

Astuti, Laksmita Widya, Yunita Sari, and Suprapto. 2023. Code-mixed sentiment analysis using transformer for twitter social media data. *International Journal of Advanced Computer Science and Applications*, 14(10).

Aulia, Nofa and Indra Budi. 2019. Hate speech detection on indonesian long text documents using machine learning approach. In *Proceedings of the 2019 5th International Conference on Computing and Artificial Intelligence*, ICCAI '19, page 164–169, Association for Computing Machinery, New York, NY, USA.

Aung, Thura, Eaint Kay Khaing Kyaw, and Ye Kyaw Thu. 2024. An empirical study of simple text augmentation on news classification for myanmar language. Preprint, Language Understanding Laboratory, Myanmar.

Aung, Thura and Pyi Hein San. 2025. Askcoviddrbot: Retrieval based tf-idf english and burmese bilingual chatbot for covid-19 domain. GitHub repository.

Buschbeck, Bianka and Miriam Exel. 2020. A parallel evaluation data set of software documentation with document structure annotation.

Cahyawijaya, Samuel, Holy Lovenia, Alham Fikri Aji, Genta Indra Winata, Bryan Wilie, Fajri Koto, Rahmad Mahendra, Christian Wibisono, Ade Romadhony, Karissa Vincentio, Jennifer Santoso, David Moeljadi, Cahya Wirawan, Frederikus Hudi, Muhammad Satrio Wicaksono, Ivan Halim Parmonangan, Ika Alfina, Ilham Firdausi Putra, Samsul Rahmadani, Yulianti Oenang, Ali Akbar Septiandri, James Jaya,

Kaustubh D. Dhole, Arie Ardiyanti Suryani, Rifki Afina Putri, Dan Su, Keith Stevens, Made Nindyatama Nityasya, Muhammad Farid Adilazuarda, Ryan Hadiwijaya, Ryandito Diandaru, Tiezheng Yu, Vito Ghifari, Wenliang Dai, Yan Xu, Dyah Damapuspita, Haryo Akbarianto Wibowo, Cuk Tho, Ichwanul Muslim Karo Karo, Tirana Fatyanosa, Ziwei Ji, Graham Neubig, Timothy Baldwin, Sebastian Ruder, Pascale Fung, Herry Sujaini, Sakriani Sakti, and Ayu Purwarianti. 2023. Nusacrowd: Open source initiative for indonesian NLP resources. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13745–13818, Association for Computational Linguistics.

Cahyawijaya, Samuel, Holy Lovenia, Joel Ruben Antony Moniz, Tack Hwa Wong, Mohammad Rifqi Farhansyah, Thant Thiri Maung, Frederikus Hudi, David Anugraha, Muhammad Ravi Shulthan Habibi, Muhammad Reza Qorib, Amit Agarwal, Joseph Marvin Imperial, Hitesh Laxmichand Patel, Vicky Feliren, Bahrul Ilmi Nasution, Manuel Antonio Rufino, Genta Indra Winata, Rian Adam Rajagede, Carlos Rafael Catalan, Mohamed Fazli Imam, Priyaranjan Pattnayak, Salsabila Zahirah Pranida, Kevin Pratama, Yeshil Bangera, Adisai Na-Thalang, Patricia Nicole Monderin, Yueqi Song, Christian Simon, Lynnette Hui Xian Ng, Richardy Lobo' Sapan, Taki Hasan Rafi, Bin Wang, Supryadi, Kanyakorn Veerakanjana, Piyalitt Ittichaiwong, Matthew Theodore Roque, Karissa Vincentio, Takdanai Kreangphet, Phakphum Artkaew, Kadek Hendrawan Palgunadi, Yanzhi Yu, Rochana Prih Hastuti, William Nixon, Mithil Bangera, Adrian Xuan Wei Lim, Aye Hninn Khine, Hanif Muhammad Zhafran, Teddy Ferdinan, Audra Aurora Izzani, Ayushman Singh, Evan, Jauza Akbar Krito, Michael Anugraha, Fenal Ashokbhai Ilasariya, Haochen Li, John Amadeo Daniswara, Filbert Aurelian Tjiaranata, Eryawan Presma Yulianrifat, Can Udomcharoenchaikit, Fadil Risdian Ansori, Mahardika Krisna Ihsani, Giang Nguyen, Anab Maulana Barik, Dan John Velasco, Rifo Ahmad Genadi, Saptarshi Saha, Chengwei Wei, Isaiah Flores, Kenneth Ko Han Chen, Anjela Gail Santos, Wan Shen Lim, Kaung Si Phyo, Tim Santos, Meisyarah Dwiastuti, Jiayun Luo, Jan Christian Blaise Cruz, Ming Shan Hee, Ikhlasul Akmal Hanif, M. Alif Al Hakim, Muhammad Rizky Sya'ban, Kun Kerdthaisong, Lester James V. Miranda, Fajri Koto, Tirana Noor Fatyanosa, Alham Fikri Aji, Jostin Jerico Rosal, Jun Kevin, Robert Wijaya, Onno P. Kampman, Ruochen Zhang, Börje F. Karlsson, and Peerat Limkonchotiwat. 2025. Crowdsource, crawl, or generate? creating sea-vl, a multicultural vision-language dataset for southeast asia.

Cahyawijaya, Samuel, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kuncoro, Sebastian Ruder, Zhi Yuan Lim, Syafri Bahar, Masayu Khodra, Ayu Purwarianti, and Pascale Fung. 2021. IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8875–8898, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic.

Catapang, Jasper Kyle and Moses Visperas. 2023. Emotion-based morality in Tagalog and English scenarios (EMoTES-3K): A parallel corpus for explaining (im)morality of actions. In *Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities and 8th International Workshop on Computational Linguistics for Uralic Languages*, pages 1–6, Association for Computational Linguistics, Tokyo, Japan.

Cer, Daniel, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Association for Computational Linguistics, Vancouver, Canada.

Chandra, Andreas. 2020. Indonesian news dataset. Online. Accessed: 2024-02-13.

Chen, Jianlv, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation.

Chen, Xi, Ali Zeynali, Chico Camargo, Fabian Flöck, Devin Gaffney, Przemyslaw Grabowicz, Scott A. Hale, David Jurgens, and Mattia Samory. 2022. SemEval-2022 task 8: Multilingual news article similarity. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1094–1106, Association for Computational Linguistics, Seattle, United States.

Chiaranaipanich, Jirat, Naiyarat Hanmatheekuna, Jitkapat Sawatphol, Krittamate Tiankanon, Jiramet Kinchagawat, Amrest Chinkamol, Parinthapat Pengpun, Piyalitt Ittichaiwong, and Peerat Limkonchotiwat. 2024. Can general-purpose large language models generalize to english-thai machine translation ?

Chrismanto, Antonius Rachmat, Anny Kartika Sari, and Yohanes Suyanto. 2022. Spamid-pair: A novel indonesian post–comment pairs dataset containing emoji. *International Journal of Advanced Computer Science and Applications*, 13(11).

Chuang, Yung-Sung, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljacic, Shang-Wen Li, Scott Yih, Yoon Kim, and James Glass. 2022. DiffCSE: Difference-based contrastive learning for sentence embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4207–4218,

Association for Computational Linguistics, Seattle, United States.
Ciancone, Mathieu, Imene Kerboua, Marion Schaeffer, and Wissam Siblini. 2024a. Extending the massive text embedding benchmark to french. *CoRR*, abs/2405.20468.
Ciancone, Mathieu, Imene Kerboua, Marion Schaeffer, and Wissam Siblini. 2024b. Mteb-french: Resources for french sentence embedding evaluation and analysis.
Clark, Jonathan H., Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages.
community, Tatoeba. 2021. Tatoeba: Collection of sentences and translations.
Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.
Conneau, Alexis and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*, European Language Resources Association (ELRA).
Conneau, Alexis, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Association for Computational Linguistics, Brussels, Belgium.
Cruz, Jan Christian Blaise, Jose Kristian Resabal, James Lin, Dan John Velasco, and Charibeth Cheng. 2020. Investigating the true performance of transformers in low-resource languages: A case study in automatic corpus creation. *arXiv preprint arXiv:2010.11574*.
Cruz, Jan Christian Blaise, Julianne Agatha Tan, and Charibeth Cheng. 2020. Localization of fake news detection via multitask transfer learning. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2596–2604.
cstorm125, lukkiddd. 2019. prachathai67k. `https://github.com/PyThaiNLP/prachathai-67k`.
Dang, Hoang-Quan, Duc-Duy-Anh Nguyen, and Trong-Hop Do. 2022. Multi-task solution for aspect category sentiment analysis on vietnamese datasets. In *2022 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)*, pages 404–409.
Dao, Mai Hoang, Thinh Hung Truong, and Dat Quoc Nguyen. 2021. Intent Detection and Slot Filling for Vietnamese. In *Proceedings of the 22nd Annual Conference of the International Speech Communication Association (INTERSPEECH)*.
Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.
Doddapaneni, Sumanth, Rahul Aralikatte, Gowtham Ramesh, Shreyansh Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2022. Towards leaving no indic language behind: Building monolingual corpora, benchmark and models for indic languages. *Annual Meeting of the Association for Computational Linguistics*.
Dou, Longxu, Qian Liu, Fan Zhou, Changyu Chen, Zili Wang, Ziqi Jin, Zichen Liu, Tongyao Zhu, Cunxiao Du, Penghui Yang, Haonan Wang, Jiaheng Liu, Yongchi Zhao, Xiachong Feng, Xin Mao, Man Tsung Yeung, Kunat Pipatanakul, Fajri Koto, Min Si Thu, Hynek Kydlíček, Zeyi Liu, Qunshu Lin, Sittipong Sripaisarnmongkol, Kridtaphad Sae-Khow, Nirattisai Thongchim, Taechawat Konkaew, Narong Borijindargoon, Anh Dao, Matichon Maneegard, Phakphum Artkaew, Zheng-Xin Yong, Quan Nguyen, Wannaphong Phatthiyaphaibun, Hoang H. Tran, Mike Zhang, Shiqi Chen, Tianyu Pang, Chao Du, Xinyi Wan, Wei Lu, and Min Lin. 2025. Sailor2: Sailing in south-east asia with inclusive multilingual llm. *arXiv preprint arXiv:2502.12982*.
Doxolodeo, Kerenza and Adila Alfa Krisnadhi. 2024. Ac-iquad: Automatically constructed indonesian question answering dataset by leveraging wikidata. *Language Resources and Evaluation*. Publisher Copyright: extcopyright 2024, The Author(s).
Dubey, Abhimanyu, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina

Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Enevoldsen, Kenneth C., Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzeminski, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Çagatan, Akash Kundu, Martin Bernstorff, Shitao Xiao, Akshita Sukhlecha, Bhavish Pahwa, Rafal Poswiata, Kranthi Kiran GV, Shawon Ashraf, Daniel Auras, Björn Plüster, Jan Philipp Harries, Loïc Magne, Isabelle Mohr, Mariya Hendriksen, Dawei Zhu, Hippolyte Gisserot-Boukhlef, Tom Aarsen, Jan Kostkan, Konrad Wojtasik, Taemin Lee, Marek Suppa, Crystina Zhang, Roberta Rocca, Mohammed Hamdy, Andrianos Michail, John Yang, Manuel Faysse, Aleksei Vatolin, Nandan Thakur, Manan Dey, Dipam Vasani, Pranjal Chitale, Simone Tedeschi, Nguyen Tai, Artem Snegirev, Michael Günther, Mengzhou Xia, Weijia Shi, Xing Han Lù, Jordan Clive, Gayatri Krishnakumar, Anna Maksimova, Silvan Wehrli, Maria Tikhonova, Henil Panchal, Aleksandr Abramov, Malte Ostendorff, Zheng Liu, Simon Clematide, Lester James Miranda, Alena Fenogenova, Guangyu Song, Ruqiya Bin Safi, Wen-Ding Li, Alessia Borghini, Federico Cassano, Hongjin Su, Jimmy Lin, Howard Yen, Lasse Hansen, Sara Hooker, Chenghao Xiao, Vaibhav Adlakha, Orion Weller, Siva Reddy, and Niklas Muennighoff. 2025. MMTEB: massive multilingual text embedding benchmark. *CoRR*, abs/2502.13595.

Enevoldsen, Kenneth C., Márton Kardos, Niklas Muennighoff, and Kristoffer L. Nielbo. 2024. The scandinavian embedding benchmarks: Comprehensive assessment of multilingual and monolingual text embedding. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Fe, Ridi. 2019. Indonesia sentiment analysis dataset. `https://github.com/ridife/dataset-idsa`.

Federmann, Christian, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Association for Computational Linguistics, Online.

Feng, Fangxiaoyu, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *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 878–891, Association for Computational Linguistics.

FitzGerald, Jack, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2022. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages.

Gala, Jay, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

Galinato, Valfrid, Lawrence Amores, Gino Ben Magsino, and David Rafael Sumawang. 2023. Context-based profanity detection and censorship using bidirectional encoder representations from transformers.

Gao, Tianyu, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic.

Goyal, Naman, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, and Francisco Guzmán. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 19–35.

Guntara, Tri Wahyu, Alham Fikri Aji, and Radityo Eko Prasojo. 2020. Benchmarking multidomain English-Indonesian machine translation. In *Proceedings of the 13th Workshop on Building and Using Comparable Corpora*, pages 35–43, European Language Resources Association, Marseille, France.

H"am"al"ainen, Mika, Pattama Patpong, Khalid Alnajjar, Niko Partanen, and Jack Rueter. 2021. Detecting depression in thai blog posts: a dataset and a baseline. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 20–25, Association for Computational Linguistics, Online.

Hasan, Tahmid, Abhik Bhattacharjee, Md Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703.

Hernandez Urbano Jr, Rommel, Jeffrey Uy Ajero, Angelic Legaspi Angeles, Maria Nikki Hacar Quintos, Joseph Marvin Regalado Imperial, and Ramon Llabanes Rodriguez. 2021. A bert-based hate speech classifier from transcribed online short-form videos. In *2021 5th International Conference on E-Society, E-Education and E-Technology*.

Hidayatullah, Ahmad Fathan, Siwi Cahyaningtyas, and Rheza Daffa Pamungkas. 2020. Attention-based cnn-bilstm for dialect identification on javanese text. *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, pages 317–324.

Ho, Vong Anh, Duong Huynh-Cong Nguyen, Danh Hoang Nguyen, Linh Thi-Van Pham, Duc-Vu Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2020. Emotion recognition for vietnamese social media text. In *Computational Linguistics: 16th International Conference of the Pacific Association for Computational Linguistics, PACLING 2019, Hanoi, Vietnam, October 11–13, 2019, Revised Selected Papers 16*, pages 319–333, Springer.

Htet, Aung Kyaw and Mark Dras. 2024. Myanmar xnli: Building a dataset and exploring low-resource approaches to natural language inference with myanmar. PREPRINT (Version 1) available at Research Square.

Ibrohim, Muhammad Okky and Indra Budi. 2018. A dataset and preliminaries study for abusive language detection in indonesian social media. *Procedia Computer Science*, 135:222–229. The 3rd International Conference on Computer Science and Computational Intelligence (ICCSCI 2018) : Empowering Smart Technology in Digital Era for a Better Life.

Ibrohim, Muhammad Okky and Indra Budi. 2019. Multi-label hate speech and abusive language detection in Indonesian Twitter. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 46–57, Association for Computational Linguistics, Florence, Italy.

Izzan, Ahmad, Christian Wibisono, and Ilham Firdausi Putra. 2025. Netifier: Negativity classifier. GitHub repository.

Jakarta Artificial Intelligence Research.

Jiang, Albert Q., Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Jiang, Shengyi, Sihui Fu, Nankai Lin, and Yingwen Fu. 2021a. Pre-trained models and evaluation data for the khmer language. *Tsinghua Science and Technology*.

Jiang, Shengyi, Sihui Fu, Nankai Lin, and Yingwen Fu. 2022. Pretrained models and evaluation data for the khmer language. *Tsinghua Science and Technology*, 27(4):709–718.

Jiang, Shengyi, Xiuwen Huang, Xiaonan Cai, and Nankai Lin. 2021b. Pre-trained models and evaluation data for the myanmar language. In *The 28th International Conference on Neural Information Processing*, Springer International Publishing, Cham.

Khanh, Tran, Phap Trinh, Khoa Tran, Lê Trn Ân, Luan Ha, and Kiet Nguyen. 2021. An empirical investigation of online news classification on an open-domain, large-scale and high-quality dataset in vietnamese.

Khine, A. H., K. T. Nwet, and K. M. Soe. 2017. Automatic myanmar news classification. In *15th Proceedings of International Conference on Computer Applications*, pages 401–408.

Kiasati Desrul, Dhamir Raniah and Ade Romadhony. 2019. Abusive language detection on indonesian online news comments. In *2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, pages 320–325.

Koto, Fajri and Ikhwan Koto. 2020. Towards computational linguistics in minangkabau language: Studies on sentiment analysis and machine translation. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation (PACLIC)*, Vietnam.

Koto, Fajri, Jey Han Lau, and Timothy Baldwin. 2020. Liputan6: A large-scale indonesian dataset for text summarization. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for*

Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, pages 598–608.

Koto, Fajri, Afshin Rahimi, Jey Han Lau, and Timothy Baldwin. 2020. Indolem and indobert: A benchmark dataset and pre-trained language model for indonesian NLP. *CoRR*, abs/2011.00677.

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

Lamm, Matthew, Jennimaria Palomaki, Chris Alberti, Daniel Andor, Eunsol Choi, Livio Baldini Soares, and Michael Collins. 2020. Qed: A framework and dataset for explanations in question answering.

Laurer, Moritz, Wouter van Atteveldt, Andreu Salleras Casas, and Kasper Welbers. 2022. Less Annotating, More Classifying – Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT - NLI. *Preprint*. Publisher: Open Science Framework.

Lee, Chankyu, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nv-embed: Improved techniques for training llms as generalist embedding models.

Lewis, Patrick, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Li, Xianming and Jing Li. 2023. Angle-optimized text embeddings. *arXiv preprint arXiv:2309.12871*.

Liang, Davis, Hila Gonen, Yuning Mao, Rui Hou, Naman Goyal, Marjan Ghazvininejad, Luke Zettlemoyer, and Madian Khabsa. 2023. XLM-V: overcoming the vocabulary bottleneck in multilingual masked language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 13142–13152, Association for Computational Linguistics.

Livelo, E. D. and C. Cheng. 2018. Intelligent dengue infoveillance using gated recurrent neural learning and cross-label frequencies. In *2018 IEEE International Conference on Agents (ICA)*, pages 2–7.

Lovenia, Holy, Rahmad Mahendra, Salsabil Maulana Akbar, Lester James V. Miranda, Jennifer Santoso, Elyanah Aco, Akhdan Fadhilah, Jonibek Mansurov, Joseph Marvin Imperial, Onno Kampman, Joel Ruben Antony Moniz, Muhammad Ravi Shulthan Habibi, Frederikus Hudi, Jann Montalan, Ryan Hadiwijaya, Joanito Agili Lopo, William Nixon, Börje Karlsson, James Jaya, Ryandito Diandaru, Yuze Gao, Patrick Amadeus Irawan, Bin Wang, Jan Christian Blaise Cruz, Chenxi Whitehouse, Ivan Halim Parmonangan, Maria Khelli, Wenyu Zhang, Lucky Susanto, Reynard Adha Ryanda, Sonny Lazuardi Hermawan, Dan John Velasco, Muhammad Dehan Al Kautsar, Willy Fitra Hendria, Yasmin Moslem, Noah Flynn, Muhammad Farid Adilazuarda, Haochen Li, Johanes Lee, R. Damanhuri, Shuo Sun, Muhammad Reza Qorib, Amirbek Djanibekov, Wei Qi Leong, Quyet V. Do, Niklas Muennighoff, Tanrada Pansuwan, Ilham Firdausi Putra, Yan Xu, Ngee Tai Chia, Ayu Purwarianti, Sebastian Ruder, William-Chandra Tjhi, Peerat Limkonchotiwat, Alham Fikri Aji, Sedrick Keh, Genta Indra Winata, Ruochen Zhang, Fajri Koto, Zheng Xin Yong, and Samuel Cahyawijaya. 2024. Seacrowd: A multilingual multimodal data hub and benchmark suite for southeast asian languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 5155–5203, Association for Computational Linguistics.

Lowphansirikul, Lalita, Charin Polpanumas, Attapol T. Rutherford, and Sarana Nutanong. 2022. A large english-thai parallel corpus from the web and machine-generated text. *Lang. Resour. Evaluation*, 56(2):477–499.

Luc Phan, Luong, Phuc Huynh Pham, Kim Thi-Thanh Nguyen, Sieu Khai Huynh, Tham Thi Nguyen, Luan Thanh Nguyen, Tin Van Huynh, and Kiet Van Nguyen. 2021. Sa2sl: From aspect-based sentiment analysis to social listening system for business intelligence. In *Knowledge Science, Engineering and Management*, pages 647–658, Springer International Publishing, Cham.

Luu, Son T., Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2021. A large-scale dataset for hate speech detection on vietnamese social media texts. In *Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices*, pages 415–426, Springer International Publishing, Cham.

Maas, Andrew L., Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Association for Computational Linguistics, Portland, Oregon, USA.

Mahendra, Rahmad, Alham Fikri Aji, Samuel Louvan, Fahrurrozi Rahman, and Clara Vania. 2021. IndoNLI: A natural language inference dataset for Indonesian. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10511–10527, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic.

Mei Silviana Saputri, Rahmad Mahendra and Mirna Adriani. 2018. Emotion classification on indonesian twitter dataset. In *Proceedings of the 2018 International Conference on Asian Language*

Processing(IALP), pages 90–95, IEEE.
Min Si Thu,Khin Myat Noe. Myanmar-agriculture-1k.
Mollanorozy, Sepideh, Marc Tanti, and Malvina Nissim. 2023. Cross-lingual transfer learning with {P}ersian. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 89–95, Association for Computational Linguistics, Dubrovnik, Croatia.
Muennighoff, Niklas, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2024a. Generative representational instruction tuning. *CoRR*, abs/2402.09906.
Muennighoff, Niklas, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2024b. Generative representational instruction tuning.
Muennighoff, Niklas, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. MTEB: massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 2006–2029, Association for Computational Linguistics.
Ng, Raymond, Thanh Ngan Nguyen, Yuli Huang, Ngee Chia Tai, Wai Yi Leong, Wei Qi Leong, Xianbin Yong, Jian Gang Ngui, Yosephine Susanto, Nicholas Cheng, Hamsawardhini Rengarajan, Peerat Limkonchotiwat, Adithya Venkatadri Hulagadri, Kok Wai Teng, Yeo Yeow Tong, Bryan Siow, Wei Yi Teo, Wayne Lau, Choon Meng Tan, Brandon Ong, Zhi Hao Ong, Jann Railey Montalan, Adwin Chan, Sajeban Antonyrex, Ren Lee, Esther Choa, David Ong Tat-Wee, Bing Jie Darius Liu, William Chandra Tjhi, Erik Cambria, and Leslie Teo. 2025. Sea-lion: Southeast asian languages in one network.
Nguyen, Huyen TM, Hung V Nguyen, Quyen T Ngo, Luong X Vu, Vu Mai Tran, Bach X Ngo, and Cuong A Le. 2018a. Vlsp shared task: sentiment analysis. *Journal of Computer Science and Cybernetics*, 34(4):295–310.
Nguyen, Kiet, Son Quoc Tran, Luan Thanh Nguyen, Tin Van Huynh, Son Thanh Luu, and Ngan Luu-Thuy Nguyen. 2022. Vlsp 2021-vimrc challenge: Vietnamese machine reading comprehension. *VNU Journal of Science: Computer Science and Communication Engineering*, 38(2).
Nguyen, Kiet Van, Vu Duc Nguyen, Phu X. V. Nguyen, Tham T. H. Truong, and Ngan Luu-Thuy Nguyen. 2018b. Uit-vsfc: Vietnamese students' feedback corpus for sentiment analysis. In *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, pages 19–24.
Nguyen, Luan Thanh, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2021. Constructive and toxic speech detection for open-domain social media comments in vietnamese. In *Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices*, pages 572–583, Springer International Publishing, Cham.
Nguyen, Minh-Tien, Dac Viet Lai, Phong-Khac Do, Duc-Vu Tran, and Minh-Le Nguyen. 2016. VSoLSCSum: Building a Vietnamese sentence-comment dataset for social context summarization. In *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*, pages 38–48, The COLING 2016 Organizing Committee, Osaka, Japan.
Nguyen, Nhung Thi-Hong, Phuong Phan-Dieu Ha, Luan Thanh Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2021. Vietnamese complaint detection on e-commerce websites.
Nguyen, Phu-Vinh, Minh-Nam Tran, Long Nguyen, and Dien Dinh. 2025. Advancing vietnamese information retrieval with learning objective and benchmark.
Nhiem, Tran. 2023. Vietnamese instruction data corpus for large-scale finetuning of language models.
Nomoto, Hiroki, Kenji Okano, David Moeljadi, and Hideo Sawada. 2018. Tufs asian language parallel corpus (talpco). *24* , pages 436–439.
Nomoto, Hiroki, Kenji Okano, Sunisa Wittayapanyanon, and Junta Nomura. 2019. Interpersonal meaning annotation for asian language corpora: The case of tufs asian language parallel corpus (talpco). *25* , pages 846–849.
Ousidhoum, Nedjma, Shamsuddeen Hassan Muhammad, Mohamed Abdalla, Idris Abdulmumin, Ibrahim Said Ahmad, Sanchit Ahuja, Alham Fikri Aji, Vladimir Araujo, Abinew Ali Ayele, Pavan Baswani, Meriem Beloucif, Chris Biemann, Sofia Bourhim, Christine De Kock, Genet Shanko Dekebo, Oumaima Hourrane, Gopichand Kanumolu, Lokesh Madasu, Samuel Rutunda, Manish Shrivastava, Thamar Solorio, Nirmal Surange, Hailegnaw Getaneh Tilaye, Krishnapriya Vishnubhotla, Genta Winata, Seid Muhie Yimam, and Saif M. Mohammad. 2024a. Semrel2024: A collection of semantic textual relatedness datasets for 13 languages. In *Findings of the Association for Computational Linguistics: ACL 2024*, Association for Computational Linguistics.
Ousidhoum, Nedjma, Shamsuddeen Hassan Muhammad, Mohamed Abdalla, Idris Abdulmumin, Ibrahim Said Ahmad, Sanchit Ahuja, Alham Fikri Aji, Vladimir Araujo, Meriem Beloucif, Christine De Kock, Oumaima Hourrane, Manish Shrivastava, Thamar Solorio, Nirmal Surange, Krishnapriya Vishnubhotla, Seid Muhie Yimam, and Saif M. Mohammad. 2024b. SemEval-2024 task 1: Semantic textual relatedness for african and asian languages. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, Association for Computational Linguistics.

Park, Ryan, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. 2024. Disentangling length from quality in direct preference optimization.

Pasupa, Kitsuchart, Ponruedee Netisopakul, and Ratthawut Lertsuksakda. 2016. Sentiment analysis on thai children stories. *Artificial Life and Robotics*, 21(3):357–364.

Payoungkhamdee, Patomporn, Peerachet Porkaew, Atthasith Sinthunyathum, Phattharaphon Songphum, Witsarut Kawidam, Wichayut Loha-Udom, Prachya Boonkwan, and Vipas Sutantayawalee. 2021. Limesoda: Dataset for fake news detection in healthcare domain. In *2021 16th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, pages 1–6.

Pengpun, Parinthapat, Krittamate Tiankanon, Amrest Chinkamol, Jiramet Kinchagawat, Pitchaya Chairuengjitjaras, Pasit Supholkhan, Pubordee Aussavavirojekul, Chiraphat Boonnag, Kanyakorn Veerakanjana, Hirunkul Phimsiri, Boonthicha Sae-jia, Nattawach Sataudom, Piyalitt Ittichaiwong, and Peerat Limkonchotiwat. 2024a. On creating an English-Thai code-switched machine translation in medical domain. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6055–6073, Association for Computational Linguistics, Miami, Florida, USA.

Pengpun, Parinthapat, Can Udomcharoenchaikit, Weerayut Buaphet, and Peerat Limkonchotiwat. 2024b. Seed-free synthetic data generation framework for instruction-tuning LLMs: A case study in Thai. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 438–457, Association for Computational Linguistics, Bangkok, Thailand.

Phatthiyaphaibun, Wannaphong. 2020. Pythainlp/thai-lao-parallel-corpus: Thai lao parallel corpus v0.5.

Phatthiyaphaibun, Wannaphong. 2025. Lao news classification.

Phatthiyaphaibun, Wannaphong, Korakot Chaovavanich, Charin Polpanumas, Arthit Suriyawongkul, Lalita Lowphansirikul, Pattarawat Chormai, Peerat Limkonchotiwat, Thanathip Suntorntip, and Can Udomcharoenchaikit. 2023. PyThaiNLP: Thai natural language processing in Python. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 25–36, Empirical Methods in Natural Language Processing, Singapore, Singapore.

Pratiwi, Inggrid Yanuar Risca, Rosa Andrie Asmara, and Faisal Rahutomo. 2017. Study of hoax news detection using naïve bayes classifier in indonesian language. In *2017 11th International Conference on Information Communication Technology and System (ICTS)*, pages 73–78.

Purwarianti, Ayu and Ida Ayu Putu Ari Crisdayanti. 2019. Improving bi-lstm performance for indonesian sentiment analysis using paragraph vector. In *Proceedings of the 2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, pages 1–5, IEEE.

Putri, Rifki Afina and Alice Oh. 2022. IDK-MRC: Unanswerable questions for Indonesian machine reading comprehension. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6918–6933, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Ramesh, Gowtham, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Mahalakshmi J, Divyanshu Kakwani, Navneet Kumar, Aswin Pradeep, Srihari Nagaraj, Deepak Kumar, Vivek Raghavan, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh Shantadevi Khapra. 2022. Samanantar: The largest publicly available parallel corpora collection for 11 indic languages. *Trans. Assoc. Comput. Linguistics*, 10:145–162.

Reimers, Nils and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics.

Reimers, Nils and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics.

Riccosan and Karen Etania Saputra. 2023. Multilabel multiclass sentiment and emotion dataset from indonesian mobile application review. *Data in Brief*, 50.

Riccosan, Karen Etania Saputra, Galih Dea Pratama, and Andry Chowanda. 2022. Emotion dataset from indonesian public opinion. *Data in Brief*, 43:108465.

Rivière, Morgane, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn,

Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozinska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucinska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju-yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjösund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, and Lilly McNealus. 2024. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118.

Riza, Hammam, Michael Purwoadi, Gunarso, Teduh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thái, Rapid Sun, Vichet Chea, Khin Mar Soe, Khin Thandar Nwet, Masao Utiyama, and Chenchen Ding. 2019. Asian language treebank. In *Proceedings of O-COCOSDA*, National Institute of Information and Communication Technology (NICT), Japan.

Rizqullah, Muhammad Razif, Ayu Purwarianti, and Alham Fikri Aji. 2023. Qasina: Religious domain question answering using sirah nabawiyah. In *2023 10th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA)*, pages 1–6.

Samson Juan, Sarah, Suhaila Saee, and Fitri Suraya Mohamad. *Social Versus Physical Distancing: Analysis of Public Health Messages at the Start of COVID-19 Outbreak in Malaysia Using Natural Language Processing*.

Setya, Ken Nabila and Rahmad Mahendra. 2018. Semi-supervised textual entailment on indonesian wikipedia data. In *Proceedings of the 2018 International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*.

Si Thu, M. 2024. Burmese microbiology 1k dataset (1.1).

Singapore, AI. 2024. Sea-lion (southeast asian languages in one network): A family of large language models for southeast asia. `https://github.com/aisingapore/sealion`.

Singh, Shivalika, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Sebastian Ruder, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. 2025. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation.

Sitepu, Rohayani et al. 2024. Sentiment analysis in karonese tweet using machine learning algorithms. *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, 12(4):2482–2489.

Soğancıoğlu, Gizem, Hakime "Ozt"urk, and Arzucan "Ozg"ur. 2017. Biosses: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*, 33(14):i49–i58.

Sturua, Saba, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024a. jina-embeddings-v3: Multilingual embeddings with task lora. *CoRR*, abs/2409.10173.

Sturua, Saba, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024b. jina-embeddings-v3: Multilingual embeddings with task lora.

Suriyawongkul, Arthit, Ekapol Chuangsuwanich, Pattarawat Chormai, and Charin Polpanumas. 2019. Pythainlp/wisesight-sentiment: First release.

Susanto, Yosephine, Adithya Venkatadri Hulagadri, Jann Montalan, Jian Gang Ngui, Xian Bin Yong, Wei Qi Leong, Hamsawardhini Rengarajan, Peerat Limkonchotiwat, Yifan Mai, and William-Chandra Tjhi. 2025. SEA-HELM: southeast asian holistic evaluation of language models. *CoRR*, abs/2502.14301.

Team, Gemma. 2024. Gemma.

Thakur, Nandan, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Tho, C, Y Heryadi, L Lukas, and A Wibowo. 2021. Code-mixed sentiment analysis of indonesian language and javanese language using lexicon based approach. *Journal of Physics: Conference Series*, 1869(1):012084.

Tiedemann, Jörg. 2020. The Tatoeba Translation Challenge – Realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Association for Computational Linguistics, Online.

Trakultaweekoon, Kanokorn, Santipong Thaiprayoon, Pornpimon Palingoon, and Anocha Rugchatjaroen. 2019. The first wikipedia questions and factoid answers corpus in the thai language. In *2019 14th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, pages 1–4, IEEE.

Urailertprasert, Norawit, Peerat Limkonchotiwat, Supasorn Suwajanakorn, and Sarana Nutanong. 2024. SEA-VQA: Southeast Asian cultural context dataset for visual question answering. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 173–185, Association for Computational Linguistics, Bangkok, Thailand.

Van Dinh, Co, Son T. Luu, and Anh Gia-Tuan Nguyen. 2022. Detecting spam reviews on vietnamese e-commerce websites. In *Intelligent Information and Database Systems*, pages 595–607, Springer International Publishing, Cham.

Viriyayudhakorn, Kobkrit and Charin Polpanumas. 2021. iapp_wiki_qa_squad.

Wang, Kexin, Nils Reimers, and Iryna Gurevych. 2021. TSDAE: using transformer-based sequential denoising auto-encoder for unsupervised sentence embedding learning. *CoRR*, abs/2104.06979.

Wang, Liang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916, Association for Computational Linguistics, Bangkok, Thailand.

Wang, Liang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Wang, Wenhui, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers.

Wehrli, Silvan, Bert Arnrich, and Christopher Irrgang. 2023. German text embedding clustering benchmark. In *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)*, pages 187–201, Association for Computational Lingustics, Ingolstadt, Germany.

Weller, Orion, Benjamin Chang, Sean MacAvaney, Kyle Lo, Arman Cohan, Benjamin Van Durme, Dawn Lawrie, and Luca Soldaini. 2024. Followir: Evaluating and teaching information retrieval models to follow instructions.

William, Andika and Yunita Sari. 2020. Click-id: A novel dataset for indonesian clickbait headlines.

Winata, Genta Indra, Alham Fikri Aji, Samuel Cahyawijaya, Rahmad Mahendra, Fajri Koto, Ade Romadhony, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Pascale Fung, Timothy Baldwin, Jey Han Lau, Rico Sennrich, and Sebastian Ruder. 2023. Nusax: Multilingual parallel sentiment dataset for 10 indonesian local languages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 815–834, Association for Computational Linguistics, Dubrovnik, Croatia.

Xiao, Shitao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024a. C-pack: Packed resources for general chinese embeddings. In *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 641–649, ACM.

Xiao, Shitao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024b. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 641–649, Association for Computing Machinery, New York, NY, USA.

Yang, An, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Yulianti, Evi, Ajmal Kurnia, Mirna Adriani, and Yoppy Setyo Duto. 2021. Normalisation of indonesian-english code-mixed text and its effect on emotion classification. *International Journal of Advanced Computer Science and Applications*.

Zhang, Wenxuan, Hou Pong Chan, Yiran Zhao, Mahani Aljunied Jianyu Wang, Chaoqun Liu, Yue Deng, Zhiqiang Hu, Weiwen Xu Yew Ken Chia, Xin Li, and Lidong Bing. 2024. Seallms 3: Open foundation and chat multilingual large language models for southeast asian languages. *CoRR*.

Zhang, Xinyu, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. 2023. MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131.

Zhang, Yanzhao, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

1. Appendix

1.1 Model Links

The full model links are shown in Table A.11.

Model	Hugging Face Link
multilingual-e5-large-instruct	https://huggingface.co/intfloat/multilingual-e5-large-instruct
Qwen3-Embedding-8B	https://huggingface.co/Qwen/Qwen3-Embedding-8B
bge-multilingual-gemma2	https://huggingface.co/BAAI/bge-multilingual-gemma2
multilingual-e5-large	https://huggingface.co/intfloat/multilingual-e5-large
bge-m3	https://huggingface.co/BAAI/bge-m3
GritLM-7B	https://huggingface.co/GritLM/GritLM-7B
e5-mistral-7b-instruct	https://huggingface.co/intfloat/e5-mistral-7b-instruct
Qwen3-Embedding-0.6B	https://huggingface.co/Qwen/Qwen3-Embedding-0.6B
multilingual-mpnet-base	https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2
LaBSE	https://huggingface.co/sentence-transformers/LaBSE
multilingual-MiniLM-L12	https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2
Gemma-SEA-LION-v3-9B-IT	https://huggingface.co/aisingapore/Gemma-SEA-LION-v3-9B-IT
Sailor2-8B-Chat	https://huggingface.co/sail/Sailor2-8B-Chat

Table A.11: Models and Hugging Face links used for the evaluation.

1.2 Data Links

The complete dataset information, such as citations, languages, domains, annotation creators, and licenses, are shown in Tables A.12 and A.13.

1.3 Domains

For domains in SEA-BED benchmark, we include the following:

- **Academic:** Formal writing and research publications commonly found in scholarly journals, theses, and dissertations.
- **Blog:** Informal, conversational writings about a variety of topics published on websites or personal blogs.
- **Constructed:** Artificially created text or speech, often in experiments to target particular abilities.
- **Encyclopedic:** Structured, reference-based texts offering thorough and factual information on various topics.
- **Fiction:** Narrative writing that involves creative content, such as novels, short stories, and other storytelling forms.
- **Government:** Documents, reports, and publications officially issued by government agencies.
- **Legal:** Documents and texts concerning laws, legal processes, contracts, and legal theories.
- **Medical:** Scientific and clinical publications focused on healthcare, treatments, patient care, and medical studies.

- **News:** News articles and reports that address current events, political developments, economic trends, and other timely topics.
- **Non-fiction:** Texts grounded in real events and factual information, including biographies, essays, and documentaries.
- **Religious:** Writings concerning religious teachings, doctrines, sacred texts, and discussions on spirituality.
- **Reviews:** Analytical assessments of books, films, music, products, or services.
- **Social:** Messages and conversations shared on social media, online forums, and other digital platforms.
- **Spoken:** Spoken content such as speeches, dialogues, interviews, and recorded discussions.
- **Subtitles:** Written transcriptions or translations of spoken content from films, videos, or multimedia presentations.
- **Web:** Web-based content spanning diverse topics, often featuring hyperlinks and multimedia elements.
- **Written:** A broad category encompassing all forms of text-based communication, both print and digital.

1.4 Examples

Figures A.6 to A.14 provide examples for each task covered in SEA-BED benchmark.

Task	First set sentence	Second set sentence
Cross-lingual pairing	Paris is the most beautiful city in the world	Paris adalah kota tercantik di dunia.
Dialect pairing	Andrea Maisi đã mở tỉ số cho Ý ở phút thứ tư với một quả try.	แอมเดรย มาซี่ ได้เปิดการทำคะแนนในนาทีที่สี่ให้แก่อิตาลี.
Written-forms pairing	โรซาลีเล่าว่า "คุณต้องรู้ถึงอันตรายต่าง ๆ และดูว่าคุณพอจะทำอะไรได้บ้าง ที่ปรึกษาของฉัน ตอนที่เราดู การปะทุขนาดเล็กของภูเขาไฟเอตนา มีเศษวัตถุขนาดเล็กตกลงมา เขาจะบอกเราให้เข้าไปเก็บตัวอย่าง คุณต้องได้รับการฝึกอย่างดี อย่าวิ่ง ให้อยู่กับที่ และมองขึ้นด้านบน ถ้ามีวัตถุขนาดใหญ่ตกลงมาใส่ ก็จะได้หลบออกด้านข้าง" สำหรับผู้ที่ต้องการชมภูเขาไฟคุกรุ่น โรซาลี แนะนำให้ไป วานูอาตู มีภูเขาไฟชื่อ ยาซูร์ ซึ่งมีการปะทุขนาดเล็ก คล้ายกับดอกไม้ไฟ มีความสวยงาม และเป็นภูเขาไฟที่ไปง่าย เธอบอกว่า สามารถขับรถขึ้นไปเกือบถึงปากปล่อง จากนั้นก็มีบันไดคอนกรีต ที่สามารถเดินขึ้นไปได้ และยังมีม้านั่งให้นั่งเล่นด้วย นอกจากภูเขาไฟบนโลก เธอยังพบภูเขาไฟที่คุกรุ่น 71 ลูก บนดวงจันทร์ไอโอของดาวพฤหัสฯ ด้วย	โรซาลี โลเปส นักภูเขาไฟวิทยาของนาซาโปรดปรานการปีนภูเขาไฟที่ยังคุกรุ่นอยู่ เพื่อไปชมการปะทุขนาดเล็ก โดยเธอได้เยือนภูเขาไฟที่คุกรุ่นในทุกทวีปทั่วโลกมาแล้ว 63 ลูก

Figure A.6: Bitext mining examples.

Task	Text	Label
Language Identification	sadiissss kae pasti neng tegal yakin inyong	Jawa Ngapak
	kowe gurung ngerti betapa bahagianya dichat arek e seh wkwkwk	Jawa Timur
Sentiment	Namiss ko yung pusa ko nung bata pa ako, lagi ko sya katabi matulog and malambing.	Positive
	Sa tiktok ang pugad nila. Dapat ma ban na ang tiktok app	Negative
Topic Classification	மீண்டும் செல்வராகவன் படத்தில் நடிப்பீர்களா?	tamil-cinema
	இத்தகையோரை பாதுகாக்கவே இத்தகைய ஸ்கூட்டர் வடிவமைக்கப் பட்டுள்ளது.	business
Toxic Language Detection	Yang blm pada move on mending bsk piknik aja mumpung long weekend hehehe	Non hate speech
	Karna dia ga punya program hanya modal bacot doang	Hate speech

Figure A.7: Classification examples.

Task	Text	Cluster
Topic Clustering	ဤရာသီ၏ ရောဂါဖြစ်ပွားမှု ကနဦးလူနာများသည် ဇူလိုင်လ အနှောင်းပိုင်းတွင် ပေါ်လာကြသည်။	health
	အပင်များသည် အစာချက်ခြင်းကို နေမှတဆင့်ပြုလုပ်သည်။ အရိပ်လဲပေးပါသည်။	science/technology
	ဤစာရွက်စာတမ်းများကို ဂုဏ်ပြုရန် နိုင်ငံခြား အစိုးရများ၏ စိတ်ထက်သန်မှုမှာ ပြောင်းလဲနိုင်ပါသည်။	politics
	ဝါရှင်တန်၏ အတ္တလန်နာသရက်ရှာကို ၅-၃ ဖြင့် အနိုင်ရသော ပွဲတွင် ၂ ဂိုးသွင်းပြီး ၂ ဂိုး ဖန်တီးပေးခဲ့သည်။	sports

Figure A.8: Clustering examples.

Task	Query	Instruction	Relevant Document
Instruction Question Answering	Stellar คืออะไร	Stellar: เครือข่ายโอนเงินไร้พรมแดน เทคโนโลยีการโอนเงินระหว่างธนาคารในตอนนี้ถือว่าล้ำหน้ามาก ๆ นะครับ ทุกวันนี้เราสามารถโอนเงินจากบัญชีของเราไปยังบัญชีของธนาคารอื่น ๆ ได้อย่างสะดวก รวดเร็ว และไม่มีค่าธรรมเนียมใด ๆ ซึ่งไม่ใช่ทุกประเทศบนโลกนี้จะมีสิ่งอำนวยความสะดวกเหมือนกับประเทศไทยนะครับ ในประเทศอื่น ๆ การโอนเงินระหว่างบัญชียังมีค่าธรรมเนียม จะถูกจะแพงก็แล้วแต่ประเทศไป และยังใช้เวลาประมาณนึงอีกด้วยครับ ...	Stellar คือ เครือข่ายการโอนเงินแบบกระจายศูนย์ (decentralized) ที่มีเป้าหมายจะเป็นช่องทางการจ่ายเงินที่เร็ว ปลอดภัย ไร้พรมแดน และมีค่าธรรมเนียมที่ต่ำ ด้วยการใช้งานเทคโนโลยีบล็อกเชนทำให้ Stellar สามารถเชื่อมต่อทั้งบุคคลธรรมดากับองค์กร (เช่น ธนาคาร) และทำให้ผู้ใช้งานเหล่านี้สามารถส่งผ่านสินทรัพย์ไป-มาได้อย่างรวดเร็ว Stellar มีเป้าหมายที่จะ disrupt ระบบการจ่ายเงินที่ใช้กันอยู่ทุกวันนี้ ลองคิดถึงการโอนเงินข้ามประเทศดูทุกวันนี้การโอนเงินข้ามประเทศมีค่าธรรมเนียมการโอนที่แพง ...

Figure A.9: Instruction Retrieval examples.

Task	Text	Label
Sentiment	saran ku dan pengalaman ku , mending beli mobil niaga L300 atau canter . irit dan bandel .	[fuel (positive), machine (positive), others (neutral), part (neutral), price (neutral), service (neutral)]
	Sudah dari dulu Toyota selalu kasih produk super mahal dengan fitur pas pasan	[fuel (neutral), machine (neutral), others (neutral), part (negative), price (negative), service (neutral)]
Topic Classification	เกี่ยวกับซิมนะคะ พอดีซื้อซิมมาใหม่ไม่สามารถเปิดใช้งานได้	["report", "phone_issues"]
	เบอร์โทร โดน ระงับใช้บริการ ต้องทำอย่างไงค่ะ	["enquire", "suspend"]
Toxic Language Detection	Prabowo Sudah Kalah Menyebut Bantuan Jokowi Hanya Pencitraan Adalah Ratapan Pilu'	[Hate speech, Hate speech Individual, Hate speech Week]
	Wah bangke emang nih truk'	[Hate speech, Abusive, Hate speech Individual, Hate speech Week]

Figure A.10: Multi-label Classification examples.

Task	Sentence 1	Sentence 2	Label
Textual Entailment	Làm sao anh biết ? Tất cả đây là thông tin của họ lần nữa .	Thông tin này thuộc về họ .	Contradiction
	Conceptually kem skimming có hai kích thước cơ bản - sản phẩm và địa lý .	Sản phẩm và địa lý là những gì làm cho kem skimming làm việc .	Entailment
	Vui vẻ dành cho người lớn và trẻ em .	Vui vì chỉ có trẻ con .	Neutral

Figure A.11: Pair Classification examples.

1.5 Annotator Demographics

In this work, our collaborators helped us translate the data from English to Thai and Burmese for STS and NLI tasks. These people are Thai and Burmese undergrad and graduate students studying in Thailand, aged from 20 to 25 years old, who can speak English and their native language (Thai or Burmese). We use three Thai annotators and one Burmese annotator to create new datasets, as mentioned in Section 3.4. We also removed some examples that contain special characters that cannot be shown in Google Sheets.

Task	Query	Positive	Negative
Article Reranking	kapankah Radin Inten II dilahirkan?	Radin Inten II (Lampung, 1834 - Lampung, 5 Oktober 1858) adalah seorang pahlawan nasional Indonesia.\nNamanya diabadikan sebagai sebuah Bandara Radin Inten II dan perguruan tinggi IAIN Raden Intan di Lampung.	Akhirnya, Waleson menemukan cara lain. Ia berhasil memperalat Radin Ngerapat. Maka pengkhianatan pun terjadi. Radin Ngerapat mengundang Radin Inten II untuk mengadakan pertemuan. Dikatakannya bahwa ia ingin membicarakan bantuan yang diberikannya kepada Radin Inten II. Tanpa curiga, Radin Inten II memenuhi undangan itu. Pertemuan diadakan malam tanggal 5 Oktober 1856 di suatu tempat dekat Kunyanya. Radin Inten II ditemani oleh satu orang pengikutnya. Radin Ngerapat disertai pula oleh beberapa orang. Akan tetapi, di tempat yang cukup tersembunyi, beberapa orang serdadu Belanda sudah disiapkan untuk bertindak bila diperlukan. Radin Ngerapat mempersilahkan Radin Inten II dan pengiringnya memakan makanan yang sengaja dibawanya terlebih dahulu.

Figure A.12: Reranking examples.

Task	Query	Relevant Document
Article Retrieval	Hà Nội: Đưa vào hoạt động trạm biến áp 110kV Bắc Thành Công	Việc đầu tư dự án 'Xây dựng mới Trạm 110kV Bắc Thành Công và nhánh rẽ' sẽ góp phần giảm được tổn hao công suất và điện năng của lưới điện trong khu vực, nâng cao chất lượng điện năng.
Long Document Retrieval	มะแว้งต้นมีประโยชน์อย่างไรในเชิงสรรพคุณ?	มะแว้งต้น ประโยชน์ดีๆ สรรพคุณเด่นๆ และข้อมูลงานวิจัย\nหน้าแรก > บทความทั้งหมด > มะแว้งต้น\nชื่อสมุนไพร มะแว้งต้น\nชื่ออื่นๆ/ชื่อท้องถิ่น มะแคว้งขม, มะแคว้งดำ, มะแคว้ง (ภาคเหนือ) ,หมากแข้ง , หมากแข้งขม (ภาคอีสาน) , มะแว้ง (ภาคกลาง) , แว้งกาม (สงขลา,สุราษฎร์ธานี,ภาคใต้) , สะกั้งแค (กะเหรี่ยง-แม่ฮ่องสอน) , หมากแข้งคง (ไทยใหญ่ – แม่ฮ่องสอน , ฉาน) , เทียนเฉีย ,ซื่อเทียนเฉีย (จีนกลาง)\nชื่อวิทยาศาสตร์ Solanum indicum L. (มีหนาม) Solanum sanitwongsei (ไร้หนาม)\nชื่อพ้องทางวิทยาศาสตร์ Solanum violaceum (มีหนาม)\nชื่อสามัญ Sparrow's Brinjal , Indian nightshade\nถิ่นกำเนิดมะแว้งต้น\nมีการคาดการณ์กันว่าถิ่นกำเนิดดั้งเดิมของมะแว้งต้นนั้นอยู่ในเขตร้อนของทวีปเอเชียซึ่งอาจอยู่ในประเทศ แถบเอเชียใต้ เช่น อินเดีย บังคลาเทศ เนปาล ฯลฯ รวมถึงประเทศแถบเอเชียตะวันออกเฉียงใต้ เช่น ไทย, พม่า , ลาว ,กัมพูชา ฯลฯ ...
Question Answering	Dimana Jamie Richard Vardy lahir?	Jamie Richard Vardy (lahir dengan nama Gill; 11 January 1987) adalah pemain sepak bola Inggris yang bermain di klub Premiere League Leicester City dan tim nasional Inggris. Ia bermain sebagai striker, namun juga bisa bermain di sayap.

Figure A.13: Retrieval examples.

Task	Sentence 1	Sentence 2	Score
Multilingual STS	လူတစ်ယောက်သည် ဘေ့စ်ဘောအသင်းတွင် ရှိနေသည်။	လူတစ်ဦးသည် အသင်းတစ်သင်းတွင် ဘတ်စကတ်ဘောကစားနေသည်။	2.4
	Istilah benda hitam pertama kali diperkenalkan oleh Gustav Kirchhoff tahun 1860.	Istilah "benda hitam" pertama kali diperkenalkan oleh Gustav Robert Kirchhoff pada tahun 1862.	5
	ชายคนหนึ่งในรถสีเขียวกำลังทำความสะอาดถนนในเมือง	ชายและหญิงสามคนข้ามถนนในเมืองที่พลุกพล่าน	1.7
Cross-lingual STS	This triggered a revolution in the earth sciences.	இக் கோட்பாடு புவி அறிவியல் துறைகளில் புரட்சிகரமான மாற்றங்களை ஏற்படுத்திற்று.	4
	The up-regulation of miR-146a was also detected in cervical cancer tissues.	miR-146a ၏အသုံးအနှုန်းသည်သားအိမ်ခေါင်းကင်ဆာတွင်ထိန်းချုပ်နိုင်သည်ကိုတွေ့ရှိရသည်။	4
	A person is on a baseball team.	มีคนเล่นบาสเก็ตบอลในทีม	2.4

Figure A.14: STS examples.

1.6 Example of Our Evaluation Tool

Similar to the previous sentence embedding benchmarks (Muennighoff et al. 2023; Enevoldsen et al. 2025), the evaluation tool of SEA-BED can be simply run using Python as shown in Figure A.15. We will release all the evaluation tools, codes, results, and datasets in the final version of our paper.

```python
from seabed import SEABED
from seabed.results_to_dataframe import results_to_dataframe
from sentence_transformers import SentenceTransformer

# Define the sentence-transformers model name
model_name = "sentence-transformers/paraphrase-multilingual-mpnet-base-v2"

model = SentenceTransformer(model_name)
evaluation = SEABED(task_types=["STS", "PairClassification"])
results = evaluation.run(model, output_folder=f"results/{model_name}", batch_size=32)
results_to_dataframe(results, output_path=f"results/{model_name}")
```

Figure A.15: Example usage of the SEA-BED evaluation framework for Semantic Textual Similarity (STS) and Pair Classification tasks.

Type	Name	Languages	Domains	Sample creation	Annotations creators	License
BitextMining	ALT (Riza et al. 2019)	['ind', 'tha', ...]	['News', 'Written']	found	expert-annotated	CC BY 4.0
	BibleNLP (Akerman et al. 2023)	['ind', 'tha', 'vie', ...]	['Religious', 'Written']	found	expert-annotated	CC BY 4.0
	Flores (Goyal et al. 2022)	['ind', 'tha', 'vie', ...]	['Non-fiction', 'Encyclopaedic', 'Written']	found	human-annotated	
	Embassy (Phatthiyaphaibun 2020)	['tha', 'lao']	['Government', 'News']	found	human-annotated	CC0-1.0
	IN22Conv (Gala et al. 2023)	['tam']	['Social', 'Spoken', 'Fiction', ...]	found	expert-annotated	CC BY 4.0
	IN22Gen (Gala et al. 2023)	['tam']	['Web', 'Legal', 'Government', ...]	found	expert-annotated	CC BY 4.0
	IndoGeneral (Guntara, Aji, and Prasojo 2020)	['ind']	['General', 'Writen']	found	derived	CC BY-SA 4.0
	IndoIdentic (Gala et al. 2023)	['ind']	['News', 'Spoken', 'Web', ...]	found	derived	
	IndoNLG (Cahyawijaya et al. 2021)	['ind']	['religion']	found	derived	
	IndoNews (Guntara, Aji, and Prasojo 2020)	['ind']	['News', 'Written']	found	derived	CC BY-SA 4.0
	IndoReligious (Guntara, Aji, and Prasojo 2020)	['ind']	['Religion', 'Writen']	found	derived	CC BY-SA 4.0
	Liputan6 (Koto, Lau, and Baldwin 2020)	['ind']	['News', 'Written']	found	human-annotated	CC BY-SA 4.0
	MADLAD400 (Kudugunta et al. 2023)	['tet']	['Web']	found	derived	ODC-BY
	NTREX (Federmann, Kocmi, and Xin 2022)	['ind', 'tha', 'vie', ...]	['News', 'Written']	found	expert-annotated	CC BY-SA 4.0
	NusaxMiners (Winata et al. 2023)	['ind']	['Reviews', 'Written']	found	human-annotated	CC BY-SA 4.0
	QED (Lamm et al. 2020)	['ind', 'tha', 'vie', ...]	['Education', 'Social', 'Spoken', ...]	found	human-annotated	CC BY-SA
	SCBMTEnTh2020 (Lowphansirikul et al. 2022)	['tha']	['conversation', 'Web', 'Government', ...]	found	human-annotated	CC BY-SA 4.0
	SoftwareDocumentation (Buschbeck and Exel 2020)	['ind', 'tha', 'vie', ...]	['Web', 'Product']	found	expert-annotated	CC BY-NC 4.0
	TALPCo (Nomoto et al. 2018, 2019)	['ind', 'tha', 'vie', ...]	['Conversation', 'spoken']	found	human-annotated	CC BY-4.0
	Tatoeba (Tiedemann 2020)	['ind', 'tha', 'vie', ...]	['Written']	found	human-annotated	CC BY-2.0
	TED2020 (Reimers and Gurevych 2020)	['ind', 'tha', 'vie', ...]	['Education', 'Social', 'Spoken', ...]	found	human-annotated	CC BY–NC–ND 4.0
	ThaiGov	['tha']	['Government', 'News']	found	human-annotated	PDDL
	USEmbassy (Phatthiyaphaibun et al. 2023)	['tha']	['News']	found	derived	CC0-1.0
	VSoLSCSum (Nguyen et al. 2016)	['vie']	['Social', 'Written']	found	human-annotated	CC BY-4.0
	XLSum (Hasan et al. 2021)	['ind', 'tha', 'vie', ...]	['News', 'Written']	found	human-annotated	CC BY-NC-SA 4.0
Classification	ABUSIVE (Ibrohim and Budi 2018)	['ind']	['Social', 'Written']	found	human-annotated	CC BY-SA 4.0
	AbusiveNewsComment (Kiasati Desrul and Romadhony 2019)	['ind']	['Social', 'Web', 'News', ...]	found	human-annotated	CC BY-SA 4.0
	BookmebusReviews	['khm']	['Reviews', 'Written']	found	human-annotated	
	Clickbait (William and Sari 2020)	['ind']	['News', 'Written']	found	expert-annotated	
	CodeMixed (Tho et al. 2021)	['ind']	['Social', 'Web']	found	manual curation	CC BY 3.0
	CyberbullyingLGBT	['tha']	['Social', 'Written']	found	derived	
	Depression (H"am"al"ainen et al. 2021)	['tha']	['Social', 'Web', 'News', ...]	found	human-annotated	CC BY-NC-ND 4.0
	EMoTES3K (Catapang and Visperas 2023)	['fil']	['Morality', 'Written']	found	human-annotated	Apache license 2.0
	Emoji	['tha']	['Social', 'Written']	found	human-annotated	GPL-3.0
	EmoT (Mei Silviana Saputri and Adriani 2018)	['ind']	['Social', 'Written']	found	human-annotated	MIT
	EmotionOpinion (Riccosan et al. 2022)	['ind']	['Social', 'Written']	found	human-annotated	CC BY-SA 4.0
	EmotCMT (Yulianti et al. 2021)	['ind']	['Social', 'Written']	found	derived	MIT
	Fakenews (Cruz, Tan, and Cheng 2020)	['fil']	['News', 'Written']	found	human-annotated	
	GeneralAmy (Phatthiyaphaibun et al. 2023)	['tha']	['Social', 'Written']	found	human-annotated	CC BY 3.0
	GeneratedReviewsENTH (Lowphansirikul et al. 2022)	['tha']	['conversation', 'Web', 'Written', ...]	found	human-annotated	CC BY-SA 4.0
	GKLMIPSentiment (Jiang et al. 2021b)	['mya']	['Social', 'Web', "Written]	found	derived	
	GooglePlayReview	['ind']	['Reviews', 'Written']	found	human-annotated	CC BY 4.0
	HateSpeech (Alfina et al. 2017)	['ind']	['Social', 'Written']	found	human-annotated	
	HateSpeech	['fil']	['Social', 'Written']	found	human-annotated	Apache license 2.0
	HoaxNews (Pratiwi, Asmara, and Rahutomo 2017)	['ind']	['News', 'Written']	found	human-annotated	CC BY 4.0
	HSDNofaaulia (Aulia and Budi 2019)	['fil']	['Social', 'Written']	found	human-annotated	
	IMDB (Maas et al. 2011)	['ind']	['Reviews', 'Written']	found	human-annotated	
	Indonglish (Astuti, Sari, and Suprapto 2023)	['ind']	['Social', 'Written']	found	expert-annotated	
	JaDiIde (Hidayatullah, Cahyaningtyas, and Pamungkas 2020)	['ind']	['Social', 'Written']	found	derived	
	Karonese (Sitepu et al. 2024)	['ind']	['Social', 'Web']	found	derived	
	KhineMyanmarNews (Khine, Nwet, and Soe 2017)	['mya']	['News', 'Written']	found	derived	GPL-3.0
	Krathu500	['tha']	['Social', 'Web', 'News', ...]	found	human-annotated	
	LazadaReview	['fil']	['Reviews', 'Written']	found	derived	
	LEMSentiment (Koto et al. 2020)	['ind']	['Social', 'Review', 'Written']	found	human-annotated	CC BY-SA 4.0
	LimeSoda (Payoungkhamdee et al. 2021)	['tha']	['Healthcare', 'Written']	found	human-annotated	CC BY 4.0
	MADLAD400 (Kudugunta et al. 2023)	['tet']	['Web']	found	derived	ODC-BY
	MassiveIntent (FitzGerald et al. 2022)	['ind', 'tha', 'vie', ...]	['Spoken']	found	human-annotated	CC BY 4.0
	MassiveScenario (FitzGerald et al. 2022)	['ind', 'tha', 'vie', ...]	['Spoken']	found	human-annotated	CC BY 4.0
	Minang (Koto and Koto 2020)	['ind']	['Encyclopaedic', 'Written']	found	derived	MIT
	MultiLingualSentiment (Mollanorozy, Tanti, and Nissim 2023)	['ind', 'tha', 'vie']	['Reviews', 'Written']	found	derived	
	MurasuNews	['tam']	['News', 'Written']	found	derived	CC0
	News (Khine, Nwet, and Soe 2017)	['mya']	['News', 'Written']	found	derived	GLP-3.0
	News	['zsm']	['News', 'Written']	found	derived	
	News	['khm']	['Encyclopaedic', 'Web', 'News', ...]	found	derived	
	News	['tam']	['News', 'Written']	found	derived	CC BY-SA 4.0
	News (Phatthiyaphaibun 2025)	['lao']	['News', 'Written']	found	derived	
	NewsDataset	['ind']	['News', 'Written']	found	derived	
	NusaX (Winata et al. 2023)	['ind']	['Social', 'Economics', 'Healthcare', ...]	found	expert-annotated	CC BY-SA 4.0
	PhoATIS (Dao, Truong, and Nguyen 2021)	['vie']	['Spoken']	found	expert-annotated	
	PHElectionsSA	['fil']	['Social']	found	human-annotated	
	PHElectionsTD	['fil']	['Social']	found	human-annotated	
	Profanity (Galinato et al. 2023)	['fil']	['Social']	found	human-annotated	
	ReviewShopping (Phatthiyaphaibun et al. 2023)	['tha']	['Reviews', 'Written']	found	human-annotated	CC BY 3.0
	SIB200 (Adelani et al. 2023)	['ind', 'tha', 'vie', ...]	['News', 'Written']	found	expert-annotated	CC BY-SA 4.0
	SEATranslationeseResampled (Lovenia et al. 2024)	['ind', 'tha', 'vie', ...]	['News', 'Social', 'Culture', ...]	found	derived	Apache license 2.0
	SentEmoMobileApps (Riccosan and Saputra 2023)	['ind']	['Reviews', 'Written']	found	human-annotated	
	SentimentAnalysis (Fe 2019)	['ind']	['Social', 'Written']	found	derived	CC BY-NC-ND 4.0
	ShopeeReviews (Purwarianti and Crisdayanti 2019)	['fil']	['Social', 'Written']	found	human-annotated	MLP-2.0
	SMSA	['ind']	['Reviews', 'Written']	found	derived	MIT
	SpamidPair (Chrismanto, Sari, and Suyanto 2022)	['ind']	['Social', 'Written']	found	human-annotated	CC BY 4.0
	SpamReviews (Van Dinh, Luu, and Nguyen 2022)	['vie']	['Reviews', 'Written']	found	human-annotated	CC BY-NC 4.0
	StudentFeedback (Nguyen et al. 2018b)	['vie']	['Reviews', 'Written']	found	human-annotated	MIT
	TCAS61 (Phatthiyaphaibun et al. 2023)	['tha']	['Social', 'Written']	found	human-annotated	CC BY 3.0
	The40ThaiChildrenStories (Pasupa, Netisopakul, and Lertsuksakda 2016)	['tha']	['Encyclopaedic', 'Written']	found	human-annotated	
	ThuraMyanmarNews (Aung, Kyaw, and Thu 2024)	['mya']	['News', 'Written']	found	derived	MIT
	TiktokHatespeech (Hernandez Urbano Jr et al. 2021)	['fil']	['Social', 'Written']	found	human-annotated	CC BY-SA 4.0
	Tweets (Samson Juan, Saee, and Mohamad)	['zsm']	['Reviews', 'Written']	found	derived	
	TyphoonYolandaTweets	['fil']	['Social', 'Written']	found	human-annotated	CC BY 4.0
	UITViCTSD (Nguyen, Van Nguyen, and Nguyen 2021)	['vie']	['Social', 'Written']	found	human-annotated	
	UITViHSD (Luu, Nguyen, and Nguyen 2021)	['vie']	['Social', 'Written']	found	human-annotated	
	UITViSFD (Luc Phan et al. 2021)	['vie']	['Social', 'Written']	found	human-annotated	
	UITVION (Khanh et al. 2021)	['vie']	['Social', 'Written']	found	human-annotated	
	UITVSMEC (Ho et al. 2020)	['vie']	['Social', 'Written']	found	human-annotated	
	VaccinesTweets	['ind']	['Social', 'Written']	found	human-annotated	
	ViOCD (Nguyen et al. 2021)	['vie']	['Reviews', 'Written']	found	human-annotated	
	VLSP2016Sentiment (Nguyen et al. 2018a)	['vie']	['Reviews', 'Written']	found	human-annotated	
	WisesightSentiment (Suriyawongkul et al. 2019)	['tha']	['Social', 'News', 'Written']	found	expert-annotated	CC0-1.0
	WongnaiReviews	['tha']	['Reviews', 'Written']	found	derived	LGPL-3.0

Table A.12: The datasets included in SEA-BED (part 1).

Type	Name	Languages	Domains	Sample creation	Annotations creators	License
Clustering	EMoTES3K (Catapang and Visperas 2023)	['fil']	['Morality', 'Written']	found	human-annotated	Apache license 2.0
	MurasuNews	['tam']	['News', 'Written']	found	derived	CC0
	News (Phatthiyaphaibun 2025)	['lao']	['News', 'Written']	found	derived	
	News (Jiang et al. 2022)	['khm']	['News', 'Written']	found	derived	
	News (Chandra 2020)	['ind']	['News', 'Written']	found	derived	
	News	['tam']	['News', 'Written']	found	derived	CC BY-SA 4.0
	News (Khine, Nwet, and Soe 2017)	['mya']	['News', 'Written']	found	derived	
	SIB200 (Adelani et al. 2023)	['ind', 'tha', 'vie', ...]	['News', 'Written']	found	expert-annotated	CC BY-SA 4.0
	UITVION (Khanh et al. 2021)	['vie']	['Social', 'Written']	found	human-annotated	
	ViOCD (Nguyen et al. 2021)	['vie']	['Reviews', 'Written']	found	human-annotated	
Instruction Retrieval	AlpacaInstruct	['ind']	None	found	LM-generated	Apache license 2.0
	Vietnamese52KAlpaca (Nhiem 2023)	['vie']	None	found	LM-generated	
	WangchanThaiInstruct	['tha']	['Medical', 'Finance', 'Legal', ...]	found	human-annotated	CC BY-SA 4.0
	WangchanXSyntheticInstructThai120k (Pengpun et al. 2024b)	['tha']	['Encyclopaedic', 'Written']	found	LM-generated	MIT
Multi-label Classification	BurmesePrachathai67k (Phatthiyaphaibun et al. 2023)	['mya']	['News', 'Web', 'Written']	created	human-annotated	Apache license 2.0
	CASA (Arfinda Ilmania 2018)	['ind']	['Reviews', 'Written']	found	human-annotated	MIT
	Dengue (Livelo and Cheng 2018)	['fil']	['Social', 'Written']	found	derived	GLP-3.0
	GKLMIPNews (Jiang et al. 2021a)	['khm']	['News', 'Written']	found	derived	
	HateSpeech (Ibrohim and Budi 2019)	['ind']	['Social', 'Written']	found	human-annotated	CC BY-SA 4.0
	HoASA (A. N. Azhar and Sutiono)	['ind']	['Reviews', 'Written']	found	human-annotated	MIT
	Netifier (Izzan, Wibisono, and Putra 2025)	['ind']	['Social', 'Written']	found	human-annotated	CC BY-SA 4.0
	Prachathai67k (Phatthiyaphaibun et al. 2023)	['tha']	['News', 'Web', 'Written']	found	derived	Apache license 2.0
	TrueVoiceIntent	['tha']	['Conversation']	found	derived	
	VLSP2018SAHotel (Dang, Nguyen, and Do 2022)	['vie']	['Reviews', 'Written']	found	human-annotated	
	VLSP2018SARestaurant (Dang, Nguyen, and Do 2022)	['vie']	['Reviews', 'Written']	found	human-annotated	
Pair Classification	BurmeseXNLI (Conneau et al. 2018)	['mya']	['Non-fiction', 'Fiction', 'Government']	created	human-annotated	CC BY-NC 4.0
	IDKMRCNLI	['ind']	['Encyclopaedic', 'News', 'Written']	found		
	IndicXNLI (Aggarwal, Gupta, and Kunchukuttan 2022)	['tam']	['Non-fiction', 'Fiction', 'Government']	found	expert-annotated	CC BY-NC 4.0
	IndoNLI (Mahendra et al. 2021)	['ind']	['Encyclopaedic', 'Web', 'News', ...]	found	expert-annotated	CC BY-SA 4.0
	MultilingualNLI26lang2mil7 (Laurer et al. 2022)	['ind', 'vie']	['Non-fiction', 'Fiction', 'Government']	found	machine-translated and reviewed	
	MyXNLI (Htet and Dras 2024)	['mya']	['Non-fiction', 'Fiction', 'Government']	found	human-annotated	CC BY-NC 4.0
	NewsPHNLI (Cruz et al. 2020)	['fil']	['News', 'Written']	found	human-annotated	GPL-3.0
	PAWS	['fil']	['Web']	found	human-annotated	
	SQuADNLI	['ind']	['Encyclopaedic', 'News', 'Written']	found		
	TyDIQANLI	['ind']	['Encyclopaedic', 'News', 'Written']	found		
	WReTE (Setya and Mahendra 2018)	['tha']	['Encyclopaedic', 'Web', 'News']	found	expert-annotated	MIT
	XNLI (Conneau et al. 2018)	['tha', 'vie']	['Non-fiction', 'Fiction', 'Government']	found	expert-annotated	CC BY-NC 4.0
	XNLITranslated (Conneau et al. 2018)	['khm', 'zsm', 'lao']	['Non-fiction', 'Fiction', 'Government']	machine-translated and verified	machine-translated and reviewed	CC BY-NC 4.0
Retrieval	ACIQuAD (Doxolodeo and Krisnadhi 2024)	['ind']	['Encyclopaedic', 'Written']	found	expert-annotated	CC-BY 4.0
	Agricutlure1K (Min Si Thu,Khin Myat Noe)	['mya']	['Encyclopaedic', 'Written']	found	expert-annotated	CC BY-SA 4.0
	AskCovidDrBot (Aung and San 2025)	['mya']	['Encyclopaedic', 'Written']	found	human-annotated	MIT
	ChatGPTOpenQA	['zsm']	['Encyclopaedic', 'Written']	found	LM-generated	CC BY-NC-SA 2.0
	ContextSearch (Nguyen et al. 2025)	['tha']	['STEM', 'Humanities', 'Social Sciences', ...]	found	human-annotated	MIT
	IAppWiki (Viriyayudhakorn and Polpanumas 2021)	['tha']	['Encyclopaedic', 'Web', 'News']	found	expert-annotated	MIT
	IDKMRC (Putri and Oh 2022)	['tnd']	['Encyclopaedic', 'Written']	found	human-annotated	CC BY-SA 4.0
	IndicQA (Doddapaneni et al. 2022)	['tam']	['Web', 'Written']	machine-translated and verified	human-annotated	CC BY 4.0
	IndoNLG (Cahyawijaya et al. 2021)	['ind']	['Religion', 'Writen']	found	human-annotated	CC BY-SA 4.0
	IndoQA (Jakarta Artificial Intelligence Research)	['ind']	['Web']	found	expert-annotated	CC BY-ND 4.0
	MLDR (Chen et al. 2024)	['tha']	['Encyclopaedic', 'Written']	found	LM-generated	MIT
	MLQA (Lewis et al. 2019)	['vie']	['Encyclopaedic', 'Written']	found	human-annotated	CC BY-SA 3.0
	MIRACL (Zhang et al. 2023)	['ind', 'tha']	['Encyclopaedic', 'Written']	found	expert-annotated	Apache license 2.0
	Microbiology1K (Si Thu 2024)	['mya']	['Encyclopaedic', 'Written']	found	human-annotated	CC BY-SA 4.0
	QASiNa (Rizqullah, Purwarianti, and Aji 2023)	['ind']	['Religion', 'Writen']	found	human-annotated	MIT
	ThaiWikiQA (Trakultaweekoon et al. 2019)	['tha']	['Encyclopaedic', 'Written']	found	human-annotated	CC BY-NC-SA 3.0
	TyDiQA (Clark et al.)	['ind', 'tha']	['Encyclopaedic', 'Written']	found	human-annotated	Apache license 2.0
	ViQuAD2_0 (Nguyen et al. 2022)	['vie']	['Encyclopaedic', 'Written']	found	expert-annotated	MIT
	WangchanXLegalThaiCCLRAG (Akarajaradwong et al. 2025)	['tha']	['Legal', 'Written']	found	human-annotated	MIT
	XQuAD (Artetxe, Ruder, and Yogatama 2019)	['tha', 'vie']	['Web', 'Written']	found	human-annotated	CC BY-SA 4.0
Reranking	MIRACL (Zhang et al. 2023)	['ind', 'tha']	['Encyclopaedic', 'Written']	found	expert-annotated	Apache license 2.0
STS	Biosses (Soğancıoğlu, "Ozt"urk, and "Ozg"ur 2017)	['tha', 'mya']	['Medical']	created	human-annotated	GPL-3.0
	BiossesCrosslingual (Soğancıoğlu, "Ozt"urk, and "Ozg"ur 2017)	['tha', 'mya']	['Medical']	created	human-annotated	GPL-3.0
	IndicCrosslingual (Ramesh et al. 2022)	['tam']	[News, Non-fiction, Web, ...]	found	expert-annotated	CC0-1.0
	SemRel2024 (Ousidhoum et al. 2024a)	['ind']	['Spoken', 'Written']	found	human-annotated	
	STS17 (Cer et al. 2017)	['tha', 'mya']	['News', 'Web', 'Written']	created	human-annotated	
	STS17Crosslingual (Cer et al. 2017)	['tha', 'mya']	['News', 'Web', 'Written']	created	human-annotated	
	STS22 (Chen et al. 2022)	['tha', 'mya']	['News', 'Written']	created	human-annotated	
	STS22Crosslingual (Chen et al. 2022)	['tha', 'mya']	['News', 'Written']	created	human-annotated	
	STS24 (Ousidhoum et al. 2024b)	['tha', 'mya']	['Spoken', 'Written']	created	human-annotated	
	STS24Crosslingual (Ousidhoum et al. 2024b)	['tha', 'mya']	['Spoken', 'Written']	created	human-annotated	
	STSBenchmark (Cer et al. 2017)	['ind', 'tha', 'vie', ...]	['News', 'Web', 'Written']	machine-translated and verified	machine-translated and reviewed	CC BY-SA 4.0

Table A.13: The datasets included in SEA-BED (part 2).