Lazarus NLP at SemEval-2025 Task 11: Fine-Tuning Large Language Models for Multi-Label Emotion Classification via Sentence-Label Pairing

Wilson Wongso[1,2*], **David Samuel Setiawan**[2*], **Ananto Joyoadikusumo**[2*], **Steven Limcorn**[2*]
[1]University of New South Wales [2]LazarusNLP
[*] Equal Contribution

Abstract

Multi-label emotion classification in low-resource languages remains challenging due to limited annotated data and model adaptability. To address this, we fine-tune large language models (LLMs) using a sentence-label pairing approach, optimizing efficiency while improving classification performance. Evaluating on Sundanese, Indonesian, and Javanese, our method outperforms conventional classifier-based fine-tuning and achieves strong zero-shot cross-lingual transfer. Notably, our approach ranks first in the Sundanese subset of SemEval-2025 Task 11 Track A. Our findings demonstrate the effectiveness of LLM fine-tuning for low-resource emotion classification, underscoring the importance of tailoring adaptation strategies to specific language families in multilingual contexts. Our source code is available at: https://github.com/LazarusNLP/SemEval2025-Emotion-Analysis.

1 Introduction

Emotion recognition is a challenging and multifaceted task that plays a crucial role in natural language processing (NLP) applications, including consumer sentiment analysis (Herzig et al., 2016), healthcare (Saffar et al., 2023), and human-computer interaction (Singla et al., 2024). Despite advancements in large language models (LLMs), accurately classifying nuanced emotional expressions across diverse languages remains a significant challenge, particularly for low-resource languages. SemEval-2025 Task 11 (Muhammad et al., 2025b) addresses this challenge by providing a multilingual dataset BRIGHTER (Muhammad et al., 2025a) for three distinct sub-tasks: (A) Multi-label Emotion Detection, (B) Emotion Intensity Prediction and (C) Cross-lingual Emotion Detection. This task aims to evaluate the ability of models to identify perceived emotions in text across 28 languages, including several underrepresented ones.

In this paper, we tackled Track A (supervised multi-label emotion detection) and C (cross-lingual emotion detection). Our approach focuses on fine-tuning large language models using a novel sentence-label pairing strategy to enhance performance across both monolingual and cross-lingual tasks. We specifically target three languages spoken in Indonesia: Indonesian (`ind`), Javanese (`jav`), and Sundanese (`sun`). Track A involves assigning binary labels (0 or 1) to perceived emotions, while Track C evaluates a model's ability to transfer knowledge from one language to another without access to in-language training data.

Our methodology highlights three key contributions. First, we reformulate the multi-label classification task as a series of binary classification problems, pairing each sentence with its corresponding emotion labels. This simplifies the task into multiple single-label classifications and increases the number of training samples. Second, we leverage multilingual pre-trained LLMs to transfer cross-lingual knowledge from Sundanese (supervised training set available in Track A) to Indonesian and Javanese (test sets available in Track C), taking advantage of their linguistic similarities within the same language family. Finally, instead of using a conventional binary cross-entropy (BCE) loss with a linear classifier head, we frame the task as text generation, training the model to output "yes" or "no" as predicted labels for each emotion. Our experiments reveal the effectiveness of these strategies in improving model performance across both monolingual and cross-lingual scenarios, providing insights into addressing linguistic gaps in emotion detection tasks.

2 Related Works

Emotion Classification Approaches Early emotion classification systems primarily relied on lexicon-based approaches, which mapped words

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 763–772

to predefined emotional categories (Mohammad, 2023). While effective to some extent, these methods lacked deeper semantic understanding and were purely syntactic. As a result, machine learning models and neural networks (e.g. language models) eventually replaced rule-based systems, enabling more context-aware and flexible emotion classification.

Traditionally, multi-label emotion classification has been approached as a conventional classification problem, where a pre-trained encoder-based language model is augmented with a linear classifier and trained using BCE loss. More recent innovations have explored alternative formulations. SpanEmo (Alhuzali and Ananiadou, 2021) introduced span prediction, where models learn direct associations between text spans and emotion labels, reducing ambiguity and improving classification performance on SemEval benchmarks. Meanwhile, T5 (Raffel et al., 2020) reframed downstream tasks (e.g. classification) within a text-to-text framework, which was later extended and scaled up by FLAN (Wei et al., 2022; Chung et al., 2024) to instruction tuning, improving zero-shot generalization.

Emotion Classification in Indonesian Languages The first major benchmark for Indonesian emotion classification was the EmoT dataset (Saputri et al., 2018), later standardized in IndoNLU (Wilie et al., 2020), where IndoBERT achieved state-of-the-art performance. However, most studies have focused on Indonesian, with limited work on regional languages. For Sundanese, Putra et al. (2020) employed traditional machine learning algorithms, while Wongso et al. (2022) benchmarked Sundanese BERT models on the former's dataset. NusaWrites (Cahyawijaya et al., 2023) later expanded the EmoT dataset by translating it into various regional languages, broadening multilingual evaluation.

Language Models for Languages of Indonesia Indonesian, Javanese, and Sundanese are among the most widely available language datasets from Indonesia (Aji et al., 2022) and have been integrated into various pre-trained language models. IndoBERT (Wilie et al., 2020) was specifically designed for Indonesian, while IndoBART (Cahyawijaya et al., 2021) extended support to Javanese and Sundanese. Additionally, multilingual models such as XLM-R (Conneau et al., 2020) and mBERT (Devlin et al., 2019) have incorporated these languages into their training corpora. More recently, large language models (LLMs) such as Cendol (Cahyawijaya et al., 2024), SEA-LION (Singapore, 2024), and Sahabat-AI[1] have further expanded coverage of these languages in their pre-training.

Building on these advancements, our approach extends the emerging trend of treating classification as an instruction-tuning task. By reformulating multi-label emotion classification as a text generation problem, we leverage the capabilities of modern Indonesian LLMs for cross-lingual transfer and improve generalization across low-resource languages.

3 Multi-label Emotion Classification

Muhammad et al. (2025a) introduced BRIGHTER, an emotion recognition dataset covering 28 languages, including several low-resource ones. In this study, we focused on languages spoken in Indonesia: Indonesian, Javanese, and Sundanese, and participated in tracks that included these languages.

Briefly, Track A is a supervised multi-label emotion detection task, where given a text snippet, the goal is to predict the speaker's perceived emotions by assigning a binary label to each (0 or 1). Track B focuses on emotion intensity prediction, requiring models to predict the intensity of a given emotion on a four-level ordinal scale. However, since none of the Indonesian languages were included in Track B, we did not participate. Track C is a cross-lingual emotion detection task, where models must predict emotions in a target language without access to a corresponding training set, relying instead on labeled data from another language. We present sample instances from the dataset in Appendix A.

Since Track A includes Sundanese, we trained on its subset and applied cross-lingual transfer to Indonesian and Javanese for Track C. This approach is feasible because all three languages share the same six emotion labels (anger, disgust, fear, joy, sadness, and surprise), eliminating the need to handle unseen emotions separately. Details of each language's subset are provided in Appendix A. Moreover, these three languages share a common origin within the Malayo-Polynesian language family. Previous studies (Winata et al., 2023; Cahyawijaya et al., 2021) suggest strong cross-lingual transfer among them, which leads us to hypothesize that our approach will be effective—especially if the pre-trained language model has been exposed to these languages during pre-training.

[1] https://sahabat-ai.com/

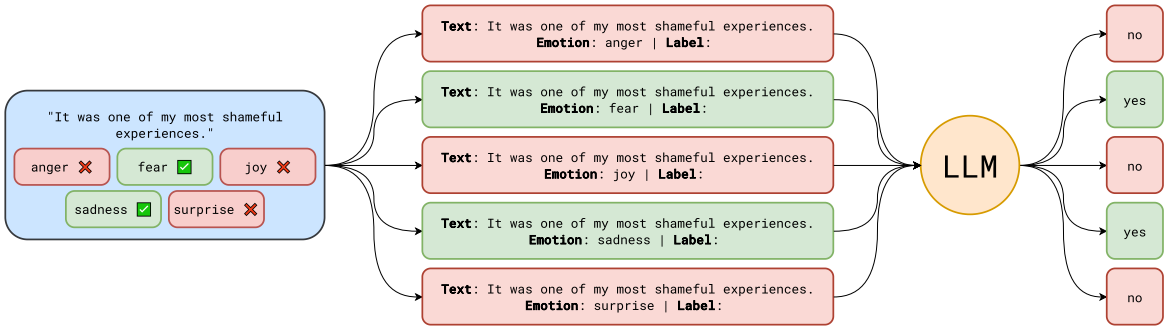

Figure 1: Our proposed LLM fine-tuning methodology for multi-label classification via sentence-label pairing.

4 System Overview

Based on our findings and problem formulation, we propose the methodology illustrated in Fig. 1. The subsequent paragraphs provide a detailed explanation of our pipeline, including data preparation, training, and inference.

4.1 Multi-Label Emotion Classification via Sentence-Label Pairing

Conventionally, with encoder-based language models such as BERT (Devlin et al., 2019), emotion classification is performed by leveraging the model's pre-trained encoder backbone and adding a linear classifier head. The model is then trained using BCE loss. The same approach can be applied to autoregressive language models, which are now commonly associated with LLM architectures. However, prior works (Radford et al., 2019; Brown et al., 2020) have demonstrated that framing tasks as text generation problems allows better utilization of an LLM's pre-trained capabilities. Building on this insight, we designed a method to adapt multi-label classification to a text generation framework.

A straightforward approach would be to concatenate the binary emotion labels into a delimited string and train the model to generate this sequence. However, we opted for a more effective strategy by reformulating the task into multiple single-label classification instances. This not only simplifies the learning process but also increases the number of training samples, which is particularly beneficial given the relatively small dataset size.

Specifically, for each text sample, we created a separate instance for each emotion label. The binary labels were then converted into natural language responses: 1 was replaced with "yes", and 0 with "no". Additional prompts (shown in Appendix B) were included to provide clearer task instructions.

4.2 Parameter-efficient LLM Supervised Fine-tuning

With the setup outlined above, we can train LLMs autoregressively for emotion classification, leveraging their natural language generation capabilities and specializing them for emotion classification through supervised fine-tuning (SFT). However, training such LLMs is computationally expensive and impractical in our resource-constrained environment.

To address these challenges and optimize both computational and time efficiency, we applied Low-Rank Adaptation (LoRA) (Hu et al., 2022), which targets only the attention layers using rank-8 matrices. We also employed QLoRA (Dettmers et al., 2023), a memory-efficient approach that quantizes model weights to 4-bit NormalFloat precision while maintaining all forward and backward passes in bfloat16, significantly reducing memory overhead. Additionally, Liger Kernels (Hsu et al., 2024) were used to further accelerate training efficiency.

4.3 Causal Inference

After SFT, we follow a standard procedure for causal inference as outlined in (Hendrycks et al., 2021). The input prompt, consisting of the text and the target emotion (e.g., happy), is passed to the LLM to generate logits. We then compare the logits for the "yes" and "no" labels, indexed by their respective token IDs. The final prediction for each emotion is the label with the higher logit score. This process is repeated for all six emotion categories.

5 Experiments

5.1 Models and Datasets

We applied our proposed methodology to two related families of LLMs: SEA-LION (Singa-

pore, 2024) and Sahabat-AI. To begin, we selected **Gemma2 9B CPT SEA-LIONv3 Base**[2] as our pre-trained LLM, which included Indonesian as part of its continued pre-training (CPT) corpus built on top of Gemma 2 (Team et al., 2024). In addition, we experimented with **Gemma2 9B CPT Sahabat-AI v1**[3], which extended SEA-LION-v3 by conducting additional CPT on English, Indonesian, Javanese, and Sundanese. Given this added exposure to the target languages, we anticipated that Sahabat-AI would deliver superior performance.

As mentioned in §3, we exclusively used Track A's Sundanese subset as our SFT dataset. Using our unpivoting method, we transformed the dataset into $924 \times 6 = 5,544$ training samples, enabling the model to learn each emotion label independently. We monitored the evaluation loss on the development subset to assess model performance during training. During inference, we applied the causal inference method to the Sundanese Track A test set, as well as the Indonesian and Javanese Track C test sets. The latter was conducted in a zero-shot and cross-lingual setting, with no additional change of prompts or modifications applied.

5.2 Implementation

The model was fine-tuned for 5 epochs with a learning rate of 2e-4 and a batch size of 32, training only the LoRA matrices. We implemented our methodology using Hugging Face Transformers (Wolf et al., 2020) and TRL (von Werra et al., 2020). All experiments were conducted on an NVIDIA L40S GPU.

5.3 Baseline

For comparison, we used NusaBERT (Wongso et al., 2025) as baseline, following the conventional approach for multi-label emotion classification explained in §4.1. A full fine-tuning was conducted with a learning rate of 1e-5, 100 epochs, early stopping with patience of 10, and a batch size of 8.

6 Results

6.1 Subtask A: Supervised Fine-tuning

We evaluated the development set results for SEA-LION-v3, Sahabat-AI, and NusaBERT fine-tuned on the Sundanese subset, with the results presented in Table 1. NusaBERT provided a solid baseline, achieving a macro F1-score of 52%. SEA-LION-v3, as expected, outperformed the conventional approach, reaching 57%. This improvement is likely due to its larger model size and enhanced capabilities, despite not being trained on Sundanese during pre-training. Nonetheless, this demonstrated the effectiveness of our proposed approach. Sahabat-AI further improved the score to 61%, which we attribute to its inclusion of Sundanese in its CPT corpus. Given its highest development score, Sahabat-AI was selected for the final testing phase on Track A and C.

Classifier	Model	Dev Score
Linear	NusaBERT Large	52%
Generative	Gemma2 9B CPT SEA-LIONv3	57%
	Gemma2 9B CPT Sahabat-AI v1	**61%**

Table 1: Macro F1-scores on the Sundanese development set for different models and training methods.

Team	Test Score (%)
SemEval Baseline (RemBERT)	37.31
PA-oneteam-1	50.72
TREDENCE AICOE	51.34
Lev Morozov	52.94
PAI	54.14
Lazarus NLP (ours)	**54.97**

Table 2: Test set macro F1-scores for Sundanese Track A, comparing our model with the top-5 leaderboard entries and the official baseline.

The test set results[4] for Sundanese Track A are shown in Table 2. Our team secured first place in the Sundanese subset, out of 38 teams on the leaderboard. However, our model's test score dropped to 54.97%, which is expected given the relatively small size of the training and development sets. Notably, most participants significantly outperformed the official baseline score of 37.31% (Muhammad et al., 2025a), where RemBERT (Chung et al., 2021) was found to be their best model under the same monolingual SFT setting.

6.2 Subtask C: Cross-lingual Transfer

We then used Track A's fine-tuned Sahabat-AI model and performed zero-shot cross-lingual transfer on the Indonesian and Javanese test sets from Track C. The results are shown in Table 3 and Table 4, respectively. Our method secured second place

[2] https://huggingface.co/aisingapore/gemma2-9b-cpt-sea-lionv3-base

[3] https://huggingface.co/GoToCompany/gemma2-9b-cpt-sahabatai-v1-base

[4] Unofficial results as of the time of writing, retrieved from the published rankings sheet.

Team	Test Score (%)
SemEval Baseline (RemBERT)	37.64
Muhammad et al. (2025a) (LaBSE)	47.50
Muhammad et al. (2025a) (Qwen2.5-72B)	57.29
deepwave	55.35
GT-NLP	58.28
Heimerdinger	60.9
Lazarus NLP (ours)	64.12
maomao	**67.24**

Table 3: Test set macro F1-scores for Indonesian Track C, comparing our model with the top-5 leaderboard entries and the official baselines.

Team	Test Score (%)
Muhammad et al. (2025a) (LaBSE)	46.24
SemEval Baseline (RemBERT)	46.38
Muhammad et al. (2025a) (Qwen2.5-72B)	**50.47**
Howard University-AI4PC	37.49
OZemi	41.26
maomao	42.20
Lazarus NLP (ours)	43.77
Heimerdinger	43.86

Table 4: Test set macro F1-scores for Javanese Track C, comparing our model with the top-5 leaderboard entries and the official baselines.

for the Indonesian subset (out of 18) and third place for the Javanese subset (out of 14).

Interestingly, our model achieved a higher score on Indonesian than Sundanese, despite the former being evaluated in a zero-shot, cross-lingual manner. We hypothesize that this may be due to the base model's stronger understanding of Indonesian, stemming from its extensive continued pre-training on a larger Indonesian corpus.

Compared to the baseline RemBERT approach provided by the organizers, our method resulted in a +26.48% improvement, further demonstrating its effectiveness. Muhammad et al. (2025a) also provided a baseline score for cross-lingual multi-label classification, where they trained on languages from the same language family–excluding the target language–and performed a cross-lingual transfer[5], reaching 47.50% with LaBSE (Feng et al., 2022). Additionally, they conducted few-shot multi-label classification, achieving 57.29% using Qwen2.5-72B (Qwen et al., 2025). Both our method and the first-place solution outperformed these approaches.

[5]In this case, training on Javanese and Sundanese (the only other two Austronesian languages), and doing a cross-lingual transfer to Indonesian.

Lang	Test Score (%)						
	Anger	Disgust	Fear	Joy	Sadness	Surprise	Overall
Zero-shot							
ind	44.55	44.56	49.24	61.33	42.44	45.29	13.51
jav	43.79	47.33	46.76	49.47	41.86	44.88	11.79
sun	47.33	48.12	48.53	53.57	43.36	48.63	14.49
Fine-tuned (ours)							
ind	79.52	75.32	76.60	81.69	81.13	64.76	64.12
jav	60.13	60.86	55.12	64.67	80.43	60.67	43.77
sun	72.02	69.75	62.07	81.80	85.41	62.71	54.97

Table 5: Test set macro F1-scores for individual emotions and overall score.

Conversely, on the Javanese subset, the organizers achieved the highest score among all participants. Qwen2.5-72B reached the top score of 50.47% via few-shot classification, outperforming our approach by +6.7%. Although not disclosed, Qwen2.5 appears to have a strong understanding of Javanese, as demonstrated by their result.

6.3 Error Analyses

Firstly, we evaluated the impact of fine-tuning by comparing the fine-tuned model with the base model under the same evaluation procedure, with results shown in Table 5. In the zero-shot setting, the Sahabat-AI model performed worst on Javanese (11.79%) and best on Sundanese (14.49%). After fine-tuning, however, the model achieved its highest F1 score on Indonesian, showing the most improvement, while Javanese performance remained the lowest. This contrasts with the pre-trained Sahabat-AI results on the SEA-HELM benchmark (Susanto et al., 2025), where the model performed best on Javanese and worst on Indonesian[6]. This suggests that the observed improvements are due not just to pre-training biases, but also to the fine-tuning data characteristics.

Secondly, to evaluate the model's performance on each emotion, we also present the per-emotion F1 scores in Table 5. For Indonesian, the lowest F1 score is for surprise (64.76%); for Javanese, it's for fear (55.12%); and for Sundanese, it's for fear (62.07%), closely followed by surprise (62.71%). To better illustrate these results, we plotted the confusion matrices in Fig. 2. Notably, the most common type of error is false negatives, which suggests a lower recall score.

Thirdly, we qualitatively evaluated five test samples with the most number of misclassified emo-

[6]Retrieved from https://hf.co/GoToCompany/gemma2-9b-cpt-sahabatai-v1-base, with scores of 60.04 on the Indonesian subset, 69.88 on the Javanese subset, and 62.44 on the Sundanese subset of the SEA-HELM benchmark.

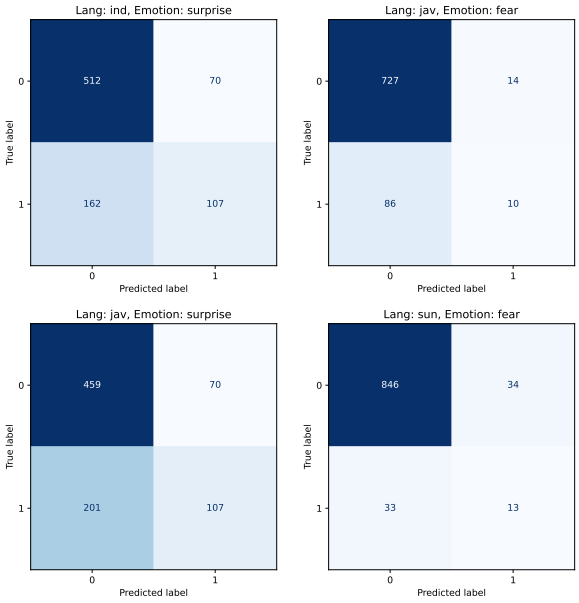

Figure 2: Confusion matrices for different languages and emotions.

tions, focusing on Indonesian texts. As shown in Table 6, we found that texts misclassified as joy often included the slang term *'wkwkwk'* (the Indonesian equivalent of *'hahaha'*). While intended sarcastically, the model struggled to grasp the true semantic meaning. Additionally, negative emotions such as anger, disgust, and fear were more likely to result in false negatives. We also identified particularly challenging samples, such as those containing Javanese slang (e.g., *'cok'*) or references to local public figures (e.g., *'Salsa Bintan'*).

7 Conclusion

In this study, we introduced a novel approach for multi-label emotion classification by fine-tuning LLMs using a sentence-label pairing method, focusing on low-resource languages of Indonesia. Our approach employed parameter- and memory-efficient techniques to optimize computational efficiency while preserving effectiveness. We trained LLMs to classify emotions from text, with Sundanese as the supervised source language and also performed cross-lingual transfer to Indonesian and Javanese. Our results demonstrated that our method outperformed conventional fine-tuning approaches in the supervised fine-tuning subtask. Additionally, our method exhibited strong performance in zero-shot, cross-lingual transfer, yielding notable results on the Indonesian and Javanese test sets. These findings indicate that our approach is a viable solution for multi-label emotion classification,

Text	Labels	Predictions
hater terbesar mahasiswa, dosen, selalu menolak hasil karya kita, padahal kita ya belajar, ampun2, wkwkw *EN: The biggest haters of students are lecturers — they always reject our work, even though we're just trying to learn. Have mercy, seriously, hahaha.*	anger, fear, sadness, surprise	joy
pliss mas denn kasih tau ke istrimu untuk menutup koment ignya biar ga ada yang nyampah. sampe kapan coba komentnya diaktifkan mulu, ga tega tau. dihujat terus & dibanding-bandjngkan sama orang lain. *EN: Please, Mas Denn, tell your wife to turn off the comments on Instagram so there's no more trash-talking. How long are you going to keep the comments open? It's really sad. She keeps getting bashed and compared to others.*	anger, disgust, fear, surprise	anger, sadness
ah gak juga gue cewek dan punya 100 daftar hal yang gue benci dari cowok wkwk *EN: Ah, not necessarily — I'm a girl and I have a list of 100 things I hate about guys, hahaha.*	anger, disgust	joy, surprise
kata kata "cok" dikhusukan orang akrab broo, kalo belum akrab jangan kek gitu bahaya soalnya. *EN: The word "cok" is reserved for close friends, bro. If you're not close, don't use it like that — it can be dangerous.*	anger, disgust, surprise	disgust, fear, joy
gendre lagunya gak cocok buat salsa bintan, jadi gak greget dan seru lagi liatnya salsa bintan, *EN: The music genre doesn't suit Salsa Bintan, so it's not exciting or fun to watch Salsa Bintan anymore.*	anger, disgust, sadness, surprise	N/A

Table 6: Test set samples from the Indonesian Track C with the highest number of misclassifications, accompanied by English translations for clarity.

particularly in low-resource settings. However, further research is needed to refine these methods and evaluate their real-world applicability.

Limitations

Our study is limited by the scope of the BRIGHTER dataset (Muhammad et al., 2025a), which focuses on Indonesian, Javanese, and Sundanese. While these languages are part of the greater Austronesian family and are among the most widely available language data in Indonesia, they belong to the smaller Central Malayo-Polynesian group (Aji et al., 2022). As a result, our findings may not be directly applicable to all languages of Indonesia nor to the broader Austronesian language family. Additionally, we did not conduct extensive hyperparameter optimization or ablation studies, limiting the potential for fine-tuning the model to achieve optimal performance across different settings and tasks.

References

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

Hassan Alhuzali and Sophia Ananiadou. 2021. SpanEmo: Casting multi-label emotion classification as span-prediction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1573–1584, Online. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Dea Adhista, Emmanuel Dave, Sarah Oktavianti, Salsabil Akbar, Jhonson Lee, Nuur Shadieq, Tjeng Wawan Cenggoro, Hanung Linuwih, Bryan Wilie, Galih Muridan, Genta Winata, David Moeljadi, Alham Fikri Aji, Ayu Purwarianti, and Pascale Fung. 2023. NusaWrites: Constructing high-quality corpora for underrepresented and extremely low-resource languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 921–945, Nusa Dua, Bali. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Rifki Putri, Wawan Cenggoro, Jhonson Lee, Salsabil Akbar, Emmanuel Dave, Nuurshadieq Nuurshadieq, Muhammad Mahendra, Rr Putri, Bryan Wilie, Genta Winata, Alham Aji, Ayu Purwarianti, and Pascale Fung. 2024. Cendol: Open instruction-tuned generative large language models for Indonesian languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14899–14914, Bangkok, Thailand. Association for Computational Linguistics.

Samuel Cahyawijaya, 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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hyung Won Chung, Thibault Févry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking embedding coupling in pre-trained language models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: efficient finetuning of quantized llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Fangxiaoyu Feng, 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)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Jonathan Herzig, Guy Feigenblat, Michal Shmueli-Scheuer, David Konopnicki, and Anat Rafaeli. 2016.

Predicting customer satisfaction in customer support conversations in social media using affective features. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, pages 115–119.

Pin-Lun Hsu, Yun Dai, Vignesh Kothapalli, Qingquan Song, Shao Tang, Siyu Zhu, Steven Shimizu, Shivam Sahni, Haowen Ning, and Yanning Chen. 2024. Liger kernel: Efficient triton kernels for llm training. *arXiv preprint arXiv:2410.10989*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Saif Mohammad. 2023. Best practices in the creation and use of emotion lexicons. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1825–1836, Dubrovnik, Croatia. Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, et al. 2025a. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages. *arXiv preprint arXiv:2502.11926*.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Seid Muhie Yimam, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine De Kock, Tadesse Destaw Belay, Ibrahim Said Ahmad, Nirmal Surange, Daniela Teodorescu, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino Ali, Vladimir Araujo, Abinew Ali Ayele, Oana Ignat, Alexander Panchenko, Yi Zhou, and Saif M. Mohammad. 2025b. SemEval task 11: Bridging the gap in text-based emotion detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Vienna, Austria. Association for Computational Linguistics.

Oddy Virgantara Putra, Fathin Muhammad Wasmanson, Triana Harmini, and Shoffin Nahwa Utama. 2020. Sundanese twitter dataset for emotion classification. In *2020 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM)*, pages 391–395.

Qwen, :, An Yang, 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, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, 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. *Journal of machine learning research*, 21(140):1–67.

Alieh Hajizadeh Saffar, Tiffany Katharine Mann, and Bahadorreza Ofoghi. 2023. Textual emotion detection in health: Advances and applications. *Journal of Biomedical Informatics*, 137:104258.

Mei Silviana Saputri, Rahmad Mahendra, and Mirna Adriani. 2018. Emotion classification on indonesian twitter dataset. In *2018 International Conference on Asian Language Processing (IALP)*, pages 90–95. IEEE.

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

Chaitanya Singla, Sukhdev Singh, Preeti Sharma, Nitin Mittal, and Fikreselam Gared. 2024. Emotion recognition for human–computer interaction using high-level descriptors. *Scientific reports*, 14(1):12122.

Yosephine Susanto, Adithya Venkatadri Hulagadri, Jann Railey Montalan, Jian Gang Ngui, Xian Bin Yong, Weiqi Leong, Hamsawardhini Rengarajan, Peerat Limkonchotiwat, Yifan Mai, and William Chandra Tjhi. 2025. Sea-helm: Southeast asian holistic evaluation of language models. *Preprint*, arXiv:2502.14301.

Gemma Team, Morgane Riviere, 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 Rogozińska, 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-Plucińska, 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 Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. Trl: Transformer reinforcement learning. `https://github.com/huggingface/trl`.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Bryan Wilie, Karissa Vincentio, Genta Indra Winata, Samuel Cahyawijaya, Xiaohong Li, Zhi Yuan Lim, Sidik Soleman, Rahmad Mahendra, Pascale Fung, Syafri Bahar, et al. 2020. Indonlu: Benchmark and resources for evaluating indonesian natural language understanding. 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 843–857.

Genta Indra Winata, 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, Dubrovnik, Croatia. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wilson Wongso, Henry Lucky, and Derwin Suhartono. 2022. Pre-trained transformer-based language models for sundanese. *Journal of Big Data*, 9(1):39.

Wilson Wongso, David Samuel Setiawan, Steven Limcorn, and Ananto Joyoadikusumo. 2025. NusaBERT: Teaching IndoBERT to be multilingual and multicultural. In *Proceedings of the Second Workshop in South East Asian Language Processing*, pages 10–26, Online. Association for Computational Linguistics.

A Dataset Details

We present the dataset statistics for each language in Table 7 and provide sample dataset entries translated into English in Table 8.

Languages	Track A (#samples)			Track C (#samples)	
	train	dev	test	dev	test
Indonesian	-	-	-	156	851
Javanese	-	-	-	151	837
Sundanese	924	199	926	199	926

Table 7: Number of samples per language in Track A and Track C.

B Prompts

During fine-tuning, we format the prompt as follows: `"### Text: {text}\n### Emotion: {emotion}\n### Label: "`, where the target text that the model learns to generate is either `"yes"` or `"no"`. This structure helps the model associate

text	anger	fear	joy	sadness	surprise
Colorado, middle of nowhere.	0	1	0	0	1
This involved swimming a pretty large lake that was over my head.	0	1	0	0	0
It was one of my most shameful experiences.	0	1	0	1	0

Table 8: Data samples from the English training set from Track A. 1 represent that the emotion is perceived from the text, and 0 otherwise.

the input text with the correct emotion label, facilitating accurate predictions during the inference process.