




Mitigating Linguistic Bias between Malay and Indonesian Languages using Masked Language Models

Ferdinand Lenchau Bit^{1*}, Iman Khaleda binti Zamri^{1*}, Amzine Toushik Wasi², Taki Hasan Rafi¹, and Dong-Kyu Chae¹(✉)

¹ Department of Computer Science, Hanyang University, South Korea

² Shahjalal University of Science and Technology, Bangladesh

{ferd1214, imankhaleda, takihr, dongkyu}@hanyang.ac.kr
azmine32@student.sust.edu

*Co-first authors.

Abstract. Language models (LMs) are essential for natural language processing (NLP) tasks, but they often exhibit biases due to inadequate or imbalanced training data, particularly in multilingual settings. These biases can lead to challenges in modeling linguistically similar low-resource languages, such as Malay and Indonesian, where mutual intelligibility complicates language differentiation. Addressing these biases is critical for enhancing the performance and fairness of NLP tools for underrepresented languages. Current LMs struggle with consistency in such scenarios, often leading to language mixing or poor prediction accuracy due to insufficient data capturing subtle linguistic differences. To tackle this, we curate a novel dataset of Malay sentences infused with Indonesian intrusions by simulating mixed-language sentences through filtering and refinement. We then fine-tune a RoBERTa model on this dataset. Empirically, this model exhibits significant improvements in word-level accuracy and language consistency compared to baseline models, indicating its ability to mitigate biases effectively. We believe that our work offers a pathway to address linguistic gaps and fosters the development of more accurate and equitable NLP tools for low-resource languages in multilingual environments.

Keywords: Bias in NLP · Malay and Indonesian language models · NLP for underrepresented languages.

1 Introduction

Language Models (LMs) have revolutionized natural language processing (NLP) by enhancing capabilities in various tasks, including machine translation, sentiment analysis, and other applications [21,34]. The development of LMs tailored to specific languages, such as Malay and Indonesian, is vital for advancing NLP applications that address regional and language-specific needs [7]. However, linguistic similarities between closely related languages like Malay and Indonesian present unique challenges, particularly for monolingual models trained to respond exclusively in Malay [5,13,31,32,33]. Unintended mixing or intrusion of Indonesian terms into Malay outputs can disrupt user experiences and undermine the effectiveness of applications designed for Malay, where precision is crucial [2,12,15].

Existing LMs optimized for Malay and Indonesian, including models like MelayuBERT, IndoBERTBase, IndoBERTFineTuned, and RoBERTa, have made significant contributions to the field [7]. Despite their successes, they often produce outputs where language intrusions negatively impact the quality and consistency of generated Malay text [12,15,29]. These intrusions, where Indonesian words appear in Malay responses, present a critical obstacle for building reliable Malay-centric NLP tools. Addressing this issue is crucial to improving the linguistic and cultural accuracy of applications designed for low-resource Malay-speaking communities, an area that has not been extensively explored [11,12].

To mitigate the issue of language intrusion, this study aims to detect and categorize Indonesian terms appearing in Malay outputs and reduce errors caused by such intrusions. Using Masked Language Modeling (MLM), our study evaluates model outputs for language consistency by identifying instances where Indonesian words occur in Malay responses. We curated a dataset by filtering and refining Malay and Indonesian sentences using resources from the Leipzig Corpora Collection [17]. Standardization of Malay sentences is performed using the Malaya NLP toolkit [6], and sentence tokenization is conducted by the Natural Language Toolkit (NLTK) [10]. Masked versions of sentences are simulated, and the BERT [3] and RoBERTa [8] models are employed to predict masked terms.

Our work contributes to improving the language consistency for Malay language models and expands the understanding of language-specific biases within LMs. Our study is inspired by bias mitigation techniques, such as MBIAS framework [18]. Our bias detection technique is specific to the challenges of Malay-Indonesian language intrusion. By improving Malay-centric NLP applications, this study hopes to support the development of more linguistically and culturally accurate tools for broader adoption within the low-resource Malay-speaking community [11,12].

2 Related Works

Pre-trained Language Models. Mass use and adoption of pre-trained LMs have transformed NLP, significantly enhancing tasks such as language generation, comprehension, sentiment analysis, and trustworthiness [7,22,28,14]. BERT, a widely used model which introduced bidirectional training for understanding a language, has achieved remarkable performances in various NLP tasks [4]. Following this are the emergence of extensions of BERT, which are tailored for specific languages and domains. RoBERTa, an optimized version of BERT improved the original architecture by training for longer and with more extensive data, resulting in a more accurate language understanding [9]. In relation to this study, specialized models like MelayuBERT and IndoBERT address the linguistic nuances of Malay and Indonesian, respectively. Indonesian language has more variants including IndoBERTBase and IndoBERTFineTuned which offer better capabilities for Indonesian NLP tasks [7,28,29]. In this study, we fine-tuned RoBERTa on our dataset, resulting in the variant named RoBERTa-MalayMLMfinetuned. This model is specially tuned for detecting language intrusions.

Masked Language Modeling (MLM). MLM is a foundational technique where specific tokens in a sentence are masked, and the model is trained to predict these masked tokens using the surrounding context [4]. This method is integral to models

Table 1: Statistics of Malay and Indonesian datasets.

Setup	Name of the Dataset/Language	# Sentences
Original Datasets	Malay Sentences	30,000
	Indonesian Sentences	30,000
	Malay Words (30K)	37,602
	Malay Words (100K)	66,805
	Indonesian Words (30K)	54,330
	Indonesian Words (100K)	104,031
Filtered Sentences Per	Malay	10,790
Language	Indonesian	16,541

Table 2: Total Count of Each Word Lists. (Note that NMA: Normalized Malay Words, NIW: Normalized Indonesian Words, UMW: Unique Malay Words, UIW: Unique Indonesian Words, SW: Shared Words, FMWL: Filtered Malay Word List, FIWL: Filtered Indonesian Word List.)

Words List	NMA	NIW	UMW	UIW	SW	FMWL	FIWL
15k Common Words	N/A	N/A	9,095	10,296	4,657	13,752	14,953
30k Words List	27,372	41,785	13,573	25,605	9,858	23,431	35,463
100k Words List	44,874	76,445	21,021	47,039	16,963	37,984	64,002

like BERT and RoBERTa, which leverage MLM to learn bidirectional contextual representations [27]. In this study, MLM is employed as the primary method for fine-tuning, enabling the detection of Indonesian intrusions in Malay-generated text. Our RoBERTa-MalayMLMFinetuned model is refined using MLM on our custom dataset specifically curated to highlight instances of language intrusion, thereby enhancing its ability to differentiate between the two languages.

Language Intrusions and Bias Mitigation. Addressing language intrusion and bias is essential for maintaining language consistency and cultural accuracy in NLP applications [26,23,25,24]. Such studies are particularly essential for closely related languages such as Malay and Indonesian. While frameworks like MBIAS provided valuable insights into managing general biases, they are not fully tailored to our specific challenges. To bridge this gap, we developed custom evaluation metrics, including the Indonesian Intrusion Rate (IIR) and Malay Prediction Rate (MPR), to effectively measure and minimize language intrusions [18].

3 Method

3.1 Data Curation

For the data curation, we used sentence and word datasets in Malay and Indonesian sourced from the Wortschatz Leipzig Corpora Collection [17]. These datasets primarily consist of news sentences in both languages. Details of each dataset, including word counts, is summarized in Table 1.

Table 3: Sample Masking and Prediction Words for Malay Sentences.

Normalized Sentence
Air beras sesuai untuk semua jenis kulit wajah, tidak kira anda mempunyai masalah kulit kering atau berjerawat sekalipun.
Masked Sentence (BERT - [MASK])
Air beras sesuai untuk semua jenis kulit wajah , tidak kira anda mempunyai masalah kulit kering atau [MASK] sekalipun .
Masked Sentence (RoBERTa - <mask>)
Air beras sesuai untuk semua jenis kulit wajah, tidak kira anda mempunyai masalah kulit kering atau <mask> sekalipun .
Masked Word: berjerawat
Predicted Words:
BERT: kulit
RoBERTa: berminyak

Sentences Filtering. We selected sentences containing 8–20 words to capture sufficient linguistic complexity while avoiding overly short or lengthy structures that could complicate processing [30]. This filtering process produced the Malay and Indonesian sentence datasets, summarized in Table 1.

Normalization and Tokenization. We performed normalization and tokenization of Malay sentences using the Malaya NLP toolkit, leveraging established methods for Malay language tokenization [1,6]. For Indonesian sentences, we used NLTK, which offers superior support specifically for the Indonesian language [10].

Word Lists. From the tokenized sentences, we prepared word lists categorized as follows: Normalized Words (NW), representing the total number of words after applying normalization processes such as removing punctuation and converting to lowercase; Unique Words (UW), consisting of words exclusive to either Malay or Indonesian; Shared Words (SW), which are common to both languages; and Filtered Word Lists (FWL), containing the remaining words after applying filtering criteria. Details of these categories are summarized in Table 2.

Masked Words Dataset. We finally generated datasets by randomly masking one word in each Malay sentence in order to evaluate the algorithms’ capacity to predict missing words. To guarantee data diversity, we created three Sets (30k Words, 100k Words, and 15k Common Words), each with 10,790 sentences that contained a single masked word [20].

3.2 Base Models and Fine-tuning

Baseline Models. For our baseline models, we utilized several pre-trained models from the Hugging Face platform [19,28,29], including MelayuBERT³, IndoBERTBase⁴,

³ <https://huggingface.co/StevenLimcorn/MelayuBERT>

⁴ <https://huggingface.co/cahya/bert-base-indonesian-522M>

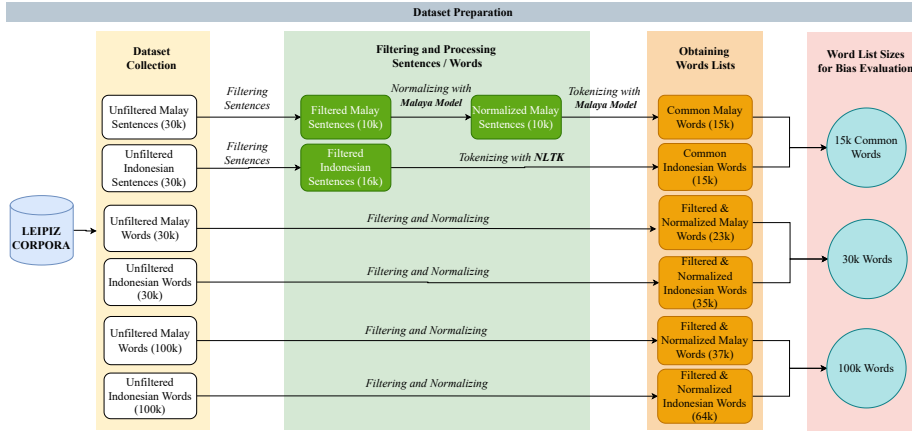


Fig. 1: Dataset Preparation Process.

IndoBERT_{FineTuned}⁵, and RoBERTa⁶. We use these models to predict masked words within the datasets. For token usage, BERT models used the [MASK] token, while RoBERTa used the <mask> token. The process of masking and predicting words from these models is further detailed in Table 3.

Fine-tuning. The fine-tuning datasets consisted of 100,790 Malay sentences sourced from the Wortschatz Leipzig Corpora Collection [17]. From this, 35,737 filtered Malay sentences are selected for fine-tuning, resulting in our RoBERTa-MalayMLM_{FineTuned} model. The hyperparameter values are summarized in Table 4. To prepare the data, 15% of the tokens in the normalized Malay sentences are randomly masked, creating three distinct masked sets [4,27].

Table 4: Summary of Key Hyperparameters for RoBERTa-MalayMLM_{FineTuned}.

Epochs	3
Batch Size	1 16
Learning Rate	5e-5
Weight Decay	0.01
Logging Steps	Every 100 steps
Evaluation Steps	Every 200 steps
Optimizer	AdamW

4 Experiments

4.1 Metrics

Language Bias Evaluation. Inspired by the MBIAS evaluation method, we evaluate the language bias using the following metrics [7,18]:

- **Indonesian Intrusion Rate (IIR):** The percentage of predicted Indonesian tokens to the total number of predicted tokens in Malay sentences.

$$\frac{\text{Number of sentences with Indonesian words}}{\text{Total number of sentences}} \times 100$$

⁵ <https://huggingface.co/racheilla/bert-base-indonesian-522M-finetuned-pemilu>

⁶ [mesolitica/roberta-base-bahasa-cased](https://huggingface.co/mesolitica/roberta-base-bahasa-cased)

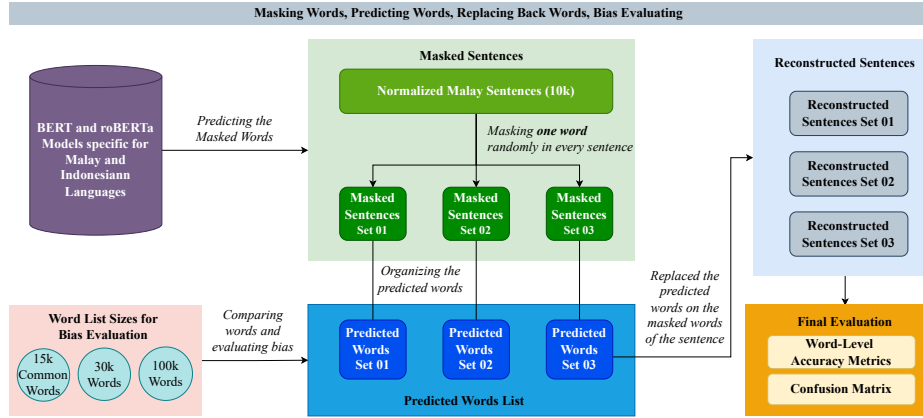


Fig. 2: Overview of our method and evaluation.

- **Unknown Intrusion Rate (UIR):** The percentage with which the model produces words that are not found in either language.

$$\frac{\text{Number of sentences with unknown words}}{\text{Total number of sentences}} \times 100$$

- **Malay Prediction Rate (MPR):** The percentage of predicted Malay words, showing how well the model performs in identifying Malay words when they are expected.

$$\frac{\text{Number of sentences with Malay words}}{\text{Total number of sentences}} \times 100$$

Then, three predefined word lists of different sizes (30k Words List, 100k Words List, and 15k Common Words) are used to compare the models' predicted word lists.

Word-level Accuracy. Word-level accuracy metrics including Precision, Recall, and F1-Score are used to assess how well the models predicted the right words [16]. To this end, we replaced the predicted words into their corresponding phrases after predicting the masked terms.

4.2 Results and Analysis

We evaluate different BERT-based models, including MelayuBERT, IndoBERTBase, IndoBERT-FineTuned, RoBERTa and **RoBERTa-MalayMLMFinetuned (Ours)**, across three datasets of varying sizes: 15k common words, 30k words, and 100k words. In Table 5, Table 6, and Table 7, we compare these models in terms of bias under different word sizes, while Table 8 presents the models' word-level accuracy with Precision, Recall, F1-score, and overall accuracy (Confusion Matrix accuracy).

According to the results, we can observe considerable performance variations among the models, particularly when measuring bias reduction and prediction accuracy. Notably, our **RoBERTa-MalayMLMFinetuned** achieved the highest average MPR with the 30k words, 100k words and 15k common words dataset while maintaining one of the lowest

Table 5: Bias Evaluation with 30k Words Size.

Model	Prediction Set	IIR ↓ (%)	MPR ↑ (%)	UIR ↓ (%)
MelayuBERT	Set 1	0.12	89.06	10.82
	Set 2	0.13	88.66	11.21
	Set 3	0.10	88.80	11.10
	Average	0.12	88.84	11.04
IndoBERTBase	Set 1	4.44	85.26	10.30
	Set 2	4.48	84.95	10.57
	Set 3	4.02	85.97	10.01
	Average	4.31	85.39	10.29
IndoBERTFineTuned	Set 1	4.46	92.34	3.20
	Set 2	4.58	92.15	3.27
	Set 3	4.59	92.30	3.11
	Average	4.54	92.27	3.19
RoBERTa	Set 1	0.32	97.36	2.32
	Set 2	0.23	97.55	2.22
	Set 3	0.26	97.58	2.16
	Average	0.27	97.50	2.23
Ours	Set 1	0.24	97.85	1.91
	Set 2	0.36	97.74	1.90
	Set 3	0.20	97.96	1.84
	Average	0.27	97.85	1.88

Table 6: Bias Evaluation with 100k Words Size.

Model	Prediction Set	IIR ↓ (%)	MPR ↑ (%)	UIR ↓ (%)
MelayuBERT	Set 1	0.05	89.28	10.68
	Set 2	0.07	88.78	11.15
	Set 3	0.03	89.00	10.97
	Average	0.05	89.02	10.93
IndoBERTBase	Set 1	3.66	86.24	10.10
	Set 2	3.50	86.14	10.36
	Set 3	3.21	86.92	9.87
	Average	3.46	86.43	10.11
IndoBERTFineTuned	Set 1	3.41	93.59	3.00
	Set 2	3.35	93.56	3.09
	Set 3	3.52	93.61	2.87
	Average	3.43	93.59	2.99
RoBERTa	Set 1	0.19	97.62	2.20
	Set 2	0.14	97.74	2.12
	Set 3	0.16	97.72	2.12
	Average	0.16	97.70	2.15
Ours	Set 1	0.18	98.12	1.71
	Set 2	0.24	98.07	1.69
	Set 3	0.12	98.20	1.68
	Average	0.18	98.13	1.69

UIR scores, indicating minimal unintended intrusions. With the best MPR and the lowest UIR across all datasets, this performance highlights the model’s ability to deliver accurate predictions and effectively mitigate biases. Moreover, it demonstrates a balanced approach to handle Malay language prediction while reducing unknown intrusions. This result may come from our additional language-specific fine-tuning to handle unknown intrusions.

The MelayuBERT model achieved an MPR between 88% and 89%, and a UIR between 10% and 11%. This suggests that MelayuBERT performs well on simpler tasks, but it has limitations in handling more complex biases. Moreover, the RoBERTa-based models demonstrate better performance at handling both unknown and Indonesian intrusions, particularly when fine-tuned with Malay language data. Consistent low IIR and UIR demonstrated by these models indicate that improved adaptation to the nuances of Malay language is possible through targeted fine-tuning. However, higher intrusion rates are observed in IndoBERTBase and IndoBERTFineTuned, highlighting the need for specialized fine-tuning in multilingual settings. Notably, the RoBERTa-MalayMLMFinetuned model effectively handles this trade-off between MPR and intrusion rates, demonstrating that language-specific optimizations benefits in mitigating biases.

Furthermore, word-level accuracy evaluations show that RoBERTa-MalayMLMFinetuned outperforms other models, achieving the highest scores across all parameters, including an accuracy of up to 42.44% on average. While MelayuBERT and IndoBERTBase consistently show lower accuracies such as 17.78% and 14.92%, respectively. Meanwhile, RoBERTa also demonstrated moderate effectiveness, with an accuracy of 36%.

RoBERTa-MalayMLMFinetuned has the potential to be utilized in tasks such as content moderation or recommendation systems that require precise Malay predictions and minimal bias. To further mitigate biases, future research should explore refining these models using different language pairs or incorporating additional linguistic nuances. In contrast, the IndoBERTFineTuned model exhibited higher variability, suggesting a

Table 7: Bias Evaluation with 15k Common Words Size.

Model	Prediction Set	IIR ↓ (%)	MPR ↑ (%)	UIR ↓ (%)
MeLayuBERT	Set 1	0.32	88.67	11.00
	Set 2	0.39	88.09	11.52
	Set 3	0.38	88.24	11.38
	Average	0.36	88.33	11.30
IndoBERTBase	Set 1	3.69	83.61	12.71
	Set 2	3.90	83.19	12.91
	Set 3	3.30	84.32	12.38
	Average	3.63	83.71	12.67
IndoBERTFineTuned	Set 1	4.40	90.06	5.53
	Set 2	4.41	90.06	5.53
	Set 3	4.22	90.50	5.28
	Average	4.34	90.21	5.45
RoBERTa	Set 1	0.66	96.54	2.80
	Set 2	0.63	96.80	2.57
	Set 3	0.52	96.93	2.55
	Average	0.60	96.76	2.64
Ours	Set 1	0.31	97.57	2.12
	Set 2	0.32	97.48	2.21
	Set 3	0.19	97.68	2.12
	Average	0.27	97.58	2.15

Table 8: Word-Level Accuracy Metrics for Each Model and Set.

Model	Pred. Set	Precision	Recall	F1-Score	Accuracy
MeLayuBERT	Set 1	0.1651	0.1762	0.1504	0.1762
	Set 2	0.1694	0.1807	0.1539	0.1807
	Set 3	0.1698	0.1766	0.1535	0.1766
	Average	0.1681	0.1778	0.1526	0.1778
IndoBERTBase	Set 1	0.1398	0.1526	0.1325	0.1526
	Set 2	0.1550	0.1654	0.1462	0.1654
	Set 3	0.1501	0.1595	0.1390	0.1595
	Average	0.1483	0.1592	0.1392	0.1592
IndoBERTFineTuned	Set 1	0.1305	0.1453	0.1225	0.1453
	Set 2	0.1456	0.1543	0.1321	0.1543
	Set 3	0.1398	0.1499	0.1257	0.1499
	Average	0.1386	0.1498	0.1268	0.1498
RoBERTa	Set 1	0.2966	0.2978	0.2814	0.3622
	Set 2	0.3012	0.2971	0.2817	0.3630
	Set 3	0.3041	0.2960	0.2814	0.3607
	Average	0.3006	0.2970	0.2815	0.3620
Ours	Set 1	0.3565	0.3518	0.3406	0.4249
	Set 2	0.3546	0.3497	0.3376	0.4264
	Set 3	0.3487	0.3446	0.3334	0.4220
	Average	0.3533	0.3487	0.3372	0.4244

strong dependence on dataset attributes; whereas the RoBERTa models performed consistently across datasets. Therefore, this performance gap suggests opportunities for improvement in token classification for Malay-Indonesian language tasks.

5 Discussion & Limitations

We believe this work has significant societal implications by not only improving the accuracy and cultural sensitivity of NLP models for low-resource languages like Malay and Indonesian but also fostering a more inclusive approach to AI development. By addressing language intrusions and biases, it directly enhances the creation of reliable, regionally tailored applications that better reflect the linguistic and cultural nuances of local communities. We believe that this improvement will lead to more effective communication tools, providing underserved communities with better access to technology and services in their native languages.

Furthermore, advances in multilingual NLP capabilities can break down language barriers, so that more people can engage with global digital platforms regardless of their language background. These advances can promote equitable access to language technologies, thereby ensuring that diverse populations are not excluded from the benefits of AI. Ultimately, this work has the potential to empower a broader global community, as well as promote social inclusion and greater digital equity for people who speak low-resource or underrepresented languages.

While our study focuses on Malay and Indonesian, the dataset size and filtering methods may limit generalizability to other dialects. Tokenization tools (NLTK for Indonesian, the Malaya toolkit for Malay) could introduce minor inconsistencies, though they do not significantly impact results. Experiments are conducted on a T4 GPU, limiting fine-tuning depth, but still sufficient for model training. Additionally, due to these computational limitations, we could not be able to go beyond RoBERTa. Future

research with more powerful GPUs could enable further optimization. Our evaluation, based on the MBIAS framework, could capture subtler biases with additional metrics. Although the masking technique’s effectiveness may vary with sentence structure and topic, it provided valuable insights, with room for refinement.

6 Conclusion

This study demonstrates the effectiveness of language-specific optimizations in enhancing BERT and RoBERTa models for Malay and Indonesian. Fine-tuning, especially with AdamW, improved prediction accuracy and reduced language intrusions. Our evaluation revealed that the proposed RoBERTa-MalayMLMFinetuned handles Malay-Indonesian data with greater precision. Additionally, our work contributes to managing multilingual bias in low-resource languages. Future research could explore larger datasets, improved tokenization, and broader language applications. Expanding the bias evaluation framework would further enhance model robustness for multilingual contexts.

Acknowledgments. This work was partly supported by (1) the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2024-00345398) and (2) the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2020-II201373,Artificial Intelligence Graduate School Program (Hanyang University)).

References

1. Bakar, J.A., Omar, K., Nasrudin, M.F., Murah, M.Z.: Tokenizer for the malay language using pattern matching. In: ICISDA. pp. 140–144 (2014)
2. BATAIS, S., WILTSHIRE, C.: Indonesian borrowing as evidence for harmonic grammar. *Journal of Linguistics* **54**(2), 231–262 (2018)
3. Devlin, J.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2019)
5. Holtermann, C., Röttger, P., Dill, T., Lauscher, A.: Evaluating the elementary multilingual capabilities of large language models with multiq (2024)
6. Husein, Z.: Natural-language-toolkit library for bahasa malaysia, powered by pytorch. <https://github.com/mesolitica/malaya> (2018)
7. Koto, F., Rahimi, A., Lau, J.H., Baldwin, T.: Indolem and indobert: A benchmark dataset and pre-trained language model for indonesian nlp. arXiv preprint arXiv:2011.00677 (2020)
8. Liu, Y.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 **364** (2019)
9. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach (2019)
10. Loper, E., Bird, S.: Nltk: The natural language toolkit. arXiv preprint (2002). <https://doi.org/10.48550/arXiv.cs/0205028>
11. Maxwell-Smith, Z., Kohler, M., Suominen, H.: Scoping natural language processing in Indonesian and Malay for education applications. In: ACL: SRW. pp. 171–228 (2022)

12. Nazri, A., et al.: Personal intelligence system unilm: Hybrid on-device small language model and server-based large language model for malay nusantara. arXiv:2410.06973 (2024)
13. Nomoto, H.: Issues surrounding the use of ChatGPT in similar languages: The case of Malay and Indonesian. In: IJCNLP-AAACL. pp. 76–82 (2023)
14. Patel, H.L., Agarwal, A., Das, A., Kumar, B., Panda, S., Pattanayak, P., Rafi, T.H., Kumar, T., Chae, D.K.: Sweeval: Do llms really swear? a benchmark for testing limits for enterprise use. In: NAACL 2025 (Industry Track) (2025)
15. Poncelas, A., Effendi, J.: Benefiting from language similarity in the multilingual MT training: Case study of Indonesian and Malaysian. In: LoResMT. pp. 84–92 (2022)
16. Powers, D.M.W.: Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation (2020)
17. Quasthoff, U.: Korpusbasierte wörterbucharbeit mit den daten des projekts deutscher wortschatz. *Linguistik online* **39**(3), 151–162 (2009)
18. Raza, S., Raval, A., Chatrath, V.: Mbias: Mitigating bias in large language models while retaining context. arXiv preprint arXiv:2405.11290 (2024)
19. Richardson, B., Wicaksana, A.: Comparison of indobert-lite and roberta in text mining for indonesian language question answering application. *IJICIC* **18**, 1719–1734 (2022)
20. Shorten, C., Khoshgoftaar, T.: Text data augmentation for deep learning. *Journal of Big Data* **8** (2021)
21. Song, L., Zhang, J., Cheng, L., Zhou, P., Zhou, T., Li, I.: Nlpbench: Evaluating large language models on solving nlp problems (2023)
22. Wang, H., Li, J., Wu, H., Hovy, E., Sun, Y.: Pre-trained language models and their applications. *Engineering* **25**, 51–65 (2023)
23. Wasi, A.T., et al.: BanglaAutoKG: Automatic Bangla knowledge graph construction with semantic neural graph filtering. In: LREC-COLING 2024. pp. 2100–2106 (May 2024)
24. Wasi, A.T., Islam, R., Islam, M.R., Rafi, T.H., Chae, D.K.: Exploring bengali religious dialect biases in large language models with evaluation perspectives. arXiv:2407.18376 (2024)
25. Wasi, A.T., Islam, R., Islam, M.R., Sadeque, F.Y., Rafi, T.H., Chae, D.K.: Dialectal bias in bengali: An evaluation of multilingual large language models across cultural variations. In: Companion Proceedings of the ACM on Web Conference 2025 (2025)
26. Wasi, A.T., Rafi, T.H., Chae, D.K.: Diaframe: A framework for understanding bengali dialects in human-ai collaborative creative writing spaces. In: CSCW. pp. 268–274 (2024)
27. Wettig, A., Gao, T., Zhong, Z., Chen, D.: Should you mask 15% in masked language modeling? In: EACL. pp. 2985–3000 (2023)
28. Wilie, B., et al.: IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. In: AACL-IJCNLP. pp. 843–857 (2020)
29. Wongso, W., Setiawan, D.S., Limcorn, S., Joyoadikusumo, A.: Nusabert: Teaching indobert to be multilingual and multicultural (2024)
30. Yaman, M.: Myparser: a malay text categorization toolkit using inference rule (2013)
31. Zampieri, M., Nakov, P., Scherrer, Y.: Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering* p. 595–612 (2020)
32. Zhang, Z., Zhao, J., Zhang, Q., Gui, T., Huang, X.: Unveiling linguistic regions in large language models. In: ACL. pp. 6228–6247 (2024)
33. Zhao, Z., Aletras, N.: Comparing explanation faithfulness between multilingual and monolingual fine-tuned language models (2024)
34. Zubiaga, A.: Natural language processing in the era of large language models. *Frontiers in Artificial Intelligence* **6** (2024)