

HATE SPEECH CLASSIFICATION IN LOW RESOURCE REGIONAL LANGUAGES USING CROSS LINGUAL TRANSFER LEARNING AND LLM BASED DATA AUGMENTATION

Vira Aditya^{1*}, Yunita Awi²

¹Informatic Engineering, Universitas Muhammadiyah Kalimantan Timur

²Informaon System, Universitas Komputer Indonesia

Email: ¹virady@gmail.com

Abstract: Hate speech on social media in Indonesia is often written in regional languages, which automatic moderation systems largely ignore because annotated data are scarce. This study builds a new annotated corpus of 6,000 social media comments in Batak Toba (BT) and Melayu Deli (MD), labelled as hate speech, abusive, or neutral by native speakers with substantial agreement (Fleiss' $\kappa = 0.71$ for BT and 0.74 for MD). Two pre-trained models, IndoBERT and XLM-RoBERTa (XLM-R), were evaluated in three scenarios: zero-shot cross-lingual transfer from an Indonesian hate speech dataset, few-shot fine-tuning, and full fine-tuning. A fourth scenario adds synthetic training data generated by a large language model (LLM) and passed through a three-stage filter covering deduplication, label consistency, and native-speaker auditing. XLM-R with full fine-tuning reached a macro-F1 of 76.1% (BT) and 80.9% (MD). Filtered LLM augmentation raised these to 79.4% and 83.1% , outperforming EDA and back-translation, whereas unfiltered synthetic data reduced performance on BT. Filtered augmentation also lowered the false positive rate on neutral comments containing ethnic or religious identity terms from 18.4% to 15.2% (BT), while unfiltered data increased it to 24.1% . These findings show that LLM-generated data can help low-resource hate speech detection, but only when quality control is applied, and they provide the first benchmark for these two languages.

Keyword: hate speech detection; low-resource languages; cross-lingual transfer; data augmentation; large language model

Abstrak: Ujaran kebencian di media sosial Indonesia sering ditulis dalam bahasa daerah, yang sebagian besar diabaikan oleh sistem moderasi otomatis karena data beranotasi sangat terbatas. Penelitian ini membangun korpus beranotasi baru berisi 6.000 komentar media sosial dalam bahasa Batak Toba (BT) dan Melayu Deli (MD), yang dilabeli sebagai ujaran kebencian, kasar, atau netral oleh penutur asli dengan tingkat kesepakatan substansial (Fleiss' $\kappa = 0,71$ untuk BT dan $0,74$ untuk MD). Dua model pralatih, IndoBERT dan XLM-RoBERTa (XLM-R), dievaluasi dalam tiga skenario: transfer lintas bahasa zero-shot dari dataset ujaran kebencian berbahasa Indonesia, fine-tuning few-shot, dan fine-tuning penuh. Skenario keempat menambahkan data latih sintetis yang dihasilkan large language model (LLM) dan disaring melalui tiga tahap, yaitu deduplikasi, konsistensi label, dan audit penutur asli. XLM-R dengan fine-tuning penuh mencapai macro-F1 $76,1\%$ (BT) dan $80,9\%$ (MD). Augmentasi LLM yang tersaring meningkatkannya menjadi $79,4\%$ dan $83,1\%$, melampaui EDA dan back-translation, sedangkan data sintetis tanpa penyaringan justru menurunkan performa pada BT. Augmentasi tersaring juga menurunkan false positive rate pada komentar netral yang memuat istilah identitas etnis atau agama dari $18,4\%$ menjadi $15,2\%$ (BT), sementara data tanpa penyaringan meningkatkannya menjadi $24,1\%$. Temuan ini menunjukkan bahwa data hasil LLM dapat membantu deteksi ujaran kebencian pada bahasa bersumber daya rendah, tetapi hanya jika disertai kendali mutu, sekaligus menyediakan tolok ukur pertama untuk kedua bahasa tersebut.

Kata kunci: deteksi ujaran kebencian; bahasa sumber daya rendah; transfer lintas bahasa; augmentasi data; large language model

INTRODUCTION

Social media has become a major arena of public discourse in Indonesia, and with it the spread of hate speech, that is, expressions that attack or demean people on the basis of ethnicity, religion, race, or other group identity. Such content can deepen social divisions and, in a country as diverse as Indonesia, carries a real risk of escalating into offline conflict. Automatic detection has therefore attracted considerable research attention, and surveys describe a mature body of methods ranging from feature-based classifiers to transformer models (Fortuna & Nunes, 2018; Davidson et al., 2022).

For the Indonesian language, several datasets and models are available. Alfina et al. (2017) released an early Indonesian hate speech dataset, and Ibrohim and Budi (2019) built a larger multi-label corpus of Indonesian tweets that distinguishes hate speech from abusive language. Pre-trained language models such as IndoBERT (Wilie et al., 2020; Koto et al., 2020) have further raised performance on Indonesian text classification. However, a great deal of everyday online conversation, particularly in regional communities, is written partly or entirely in local languages. Indonesia has more than 700 languages, and the vast majority are severely under-resourced in NLP (Aji et al., 2022). Content written in these languages can therefore pass through moderation systems that only understand Indonesian.

North Sumatra illustrates this problem. Batak Toba is spoken by a large community in and around the Lake Toba region and in the diaspora, while Melayu Deli is the Malay variety associated with the Deli area around Medan. Both appear frequently in comments on local news pages, community groups, and entertainment content, sometimes mixed with Indonesian. Recent community efforts such as NusaX (Winata et al., 2023) and NusaCrowd (Cahyawijaya et al., 2023) have begun to provide resources for Indonesian regional languages, including Batak Toba for sentiment analysis, but to the authors' knowledge no annotated hate speech corpus exists for either Batak Toba or Melayu Deli.

Two families of techniques can help when labelled data are scarce. The first is cross-lingual transfer, in which a multilingual model such as XLM-RoBERTa (Conneau et al., 2020) is fine-

tuned on a high-resource language and applied to a related low-resource one. Its effectiveness depends strongly on how closely the languages are related and on whether a few target-language examples are available (Pires et al., 2019; Lauscher et al., 2020). The second is data augmentation. Simple techniques such as easy data augmentation (EDA) (Wei & Zou, 2019) and back-translation (Sennrich et al., 2016) are widely used (Feng et al., 2021), and large language models (LLMs) now make it possible to generate realistic new examples from a handful of prompts (Brown et al., 2020). Hartvigsen et al. (2022) showed that machine-generated text can improve toxicity classifiers in English. However, LLMs are known to be weak in low-resource languages, and their output may be disfluent, mislabelled, or stereotyped, which in hate speech detection risks amplifying bias against identity groups (Sap et al., 2024).

This study addresses these gaps, and its novelty is threefold. First, it introduces the first annotated hate speech corpus for Batak Toba and Melayu Deli, labelled by native speakers. Second, it systematically compares zero-shot, few-shot, and full fine-tuning with IndoBERT and XLM-R, revealing how linguistic distance from Indonesian affects transfer. Third, it proposes a filtered LLM augmentation pipeline and evaluates the synthetic data not only by its effect on accuracy but also by its fluency, label fidelity, and effect on bias towards identity terms. The objective is to determine whether, and under what conditions, LLM-generated data can be trusted to improve hate speech detection in low-resource regional languages.

METHOD

Research Framework

The research followed the stages shown in Image 1: comment collection, anonymisation and language filtering, annotation, data splitting, model training in four scenarios, and evaluation of predictive performance and bias.

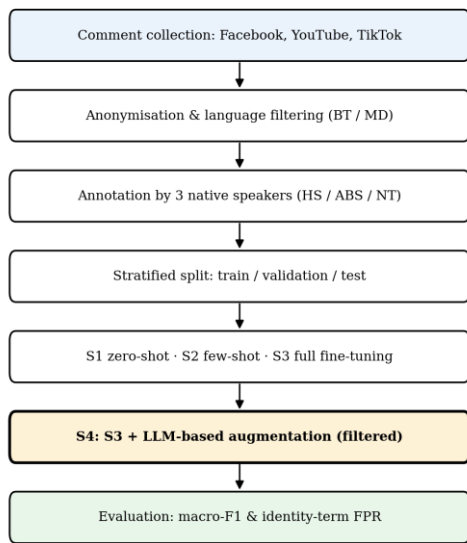


Image 1. Research framework

Data Collection

Public comments were collected from Facebook pages and groups, YouTube channels, and TikTok accounts covering local news, politics, entertainment, and community affairs in North Sumatra, posted between January 2024 and June 2026. Keyword lists in each language, including common terms related to ethnicity, religion, and politics, were used to increase the share of potentially hateful content. Comments were then filtered to retain those written mainly in Batak Toba or Melayu Deli; language identification was carried out by a keyword-based heuristic followed by manual verification by native speakers. Comments mixing the regional language with Indonesian were kept, because code-mixing is typical of real usage. All usernames, mentions, URLs, and phone numbers were replaced with placeholder tokens to protect user privacy.

Annotation

Each comment was labelled by three native speakers per language into one of three classes adapted from Ibrohim and Budi (2019). **Hate speech (HS)** targets a person or group on the basis of identity, such as ethnicity, religion, race, or gender, with insults, dehumanisation, or incitement. **Abusive (ABS)** contains profanity or insults that are not directed at group identity. **Neutral (NT)** contains neither. Annotators

received written guidelines with examples in both languages and completed a 100-comment pilot round before the main annotation. The final label was decided by majority vote, and comments without a majority were discussed and resolved by a fourth senior annotator. Inter-annotator agreement was measured using Fleiss' kappa (Fleiss, 1971), as shown in Equation (1), where $\bar{P}$ is the mean observed agreement and $\bar{P}_e$ is the agreement expected by chance.

$$\kappa = (\bar{P} - \bar{P}_e) / (1 - \bar{P}_e) \quad (1)$$

Agreement reached $\kappa = 0.71$ for Batak Toba and 0.74 for Melayu Deli, which falls in the substantial range. Because exposure to hateful content can be distressing, annotators worked in limited sessions and could skip comments. The resulting corpus is described in Table 1. Each language was split into training, validation, and test sets in a 70:10:20 ratio, stratified by class.

Table 1. Corpus statistics

Language	HS	ABS	NT	Total	κ
Batak Toba	780	960	1,260	3,000	0.71
Melayu Deli	720	900	1,380	3,000	0.74
Total	1,500	1,860	2,640	6,000	–

Models

Two pre-trained encoders were compared. IndoBERT-base (Wilie et al., 2020) was pre-trained on a large Indonesian corpus and represents a monolingual model close to the target languages. XLM-RoBERTa-base (Conneau et al., 2020) was pre-trained on text in about 100 languages, including Indonesian, Malay, Javanese, and Sundanese, and represents a multilingual model. Neither model includes Batak Toba or Melayu Deli in its pre-training data, so both must transfer knowledge from related languages. For classification, a linear layer was placed on the [CLS] representation and the whole model was fine-tuned.

Experimental Scenarios

Four scenarios were evaluated. In **S1 (zero-shot)**, each model was fine-tuned only on the Indonesian dataset of Ibrohim and Budi (2019), whose multi-label annotations were mapped into the same three classes (hate speech if the hate label was present, abusive if only the abusive label was present, otherwise neutral), and tested directly on the target languages. In **S2 (few-**

shot), the S1 model was further fine-tuned on 100 or 500 randomly sampled target-language training examples. In **S3 (full)**, the S1 model was fine-tuned on the complete target training set. In **S4 (augmented)**, the S3 setting was repeated with synthetic data added to the training set, generated by the pipeline described below. EDA (Wei & Zou, 2019) and back-translation through Indonesian using (machine translation system) (Sennrich et al., 2016) were used as baseline augmentation methods with the same number of added examples.

LLM-Based Augmentation Pipeline

The augmentation pipeline is shown in Image 2. For each language and class, an instruction-tuned LLM, (LLM name and version), was prompted with the class definition, a short description of the language variety, and five real examples randomly drawn from the training split, and asked to write new comments of the same class. The prompt explicitly asked for natural social-media style, including code-mixing, and prohibited copying the examples. A total of 3,000 candidates were generated per language.

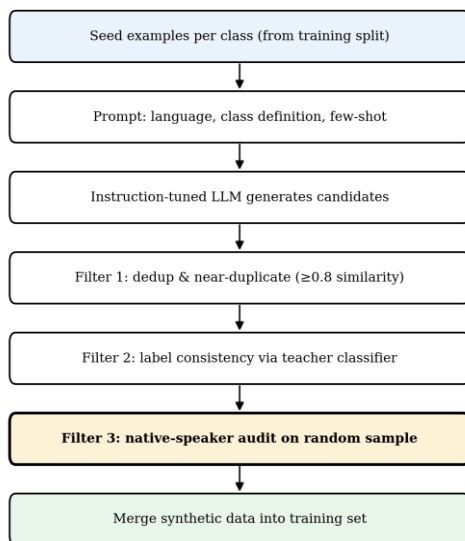


Image 2. LLM-based augmentation and filtering pipeline

Three filters were then applied. **Filter 1** removed exact duplicates, copies of seed examples, and near-duplicates with a character n-gram cosine similarity of 0.8 or higher. **Filter 2** checked label consistency by passing each candidate through the S3 classifier and

discarding those for which the predicted class differed from the intended class with a confidence above 0.9. **Filter 3** was a native-speaker audit, in which a random sample of 300 remaining candidates per language was rated for fluency on a 1–5 scale and for whether the intended label was correct; generation batches (one batch per prompt) with an audited label fidelity below 80% were discarded entirely. To measure the value of filtering, an unfiltered variant using the same number of examples was also evaluated.

Training Setup and Evaluation

All models were trained with AdamW (Loshchilov & Hutter, 2019), a learning rate of 2×10^{-5} , batch size 16, maximum sequence length 128 tokens, and up to 10 epochs with early stopping on validation macro-F1. Each experiment was repeated with five random seeds, and the mean is reported. Because the classes are imbalanced, the main metric was macro-F1 (Equation 2), the unweighted mean of the per-class F1-scores.

$$\text{Macro-F1} = (1/3) \cdot (F1_{HS} + F1_{ABS} + F1_{NT}) \quad (2)$$

To measure bias, a separate set of 150 neutral comments per language that mention ethnic or religious identity terms (for example, references to ethnic groups or places of worship in non-hateful contexts) was compiled. The identity-term false positive rate (FPR_{id}) in Equation (3) is the proportion of these neutral comments wrongly classified as hate speech or abusive, where FP_{id} is the number of such errors and TN_{id} the number correctly classified as neutral. A lower value indicates that the model is less likely to flag a comment simply because it mentions a group.

$$FPR_{id} = FP_{id} / (FP_{id} + TN_{id}) \quad (3)$$

Differences between S3 and S4 were tested with McNemar's test (McNemar, 1947) at a significance level of 0.05.

RESULT AND DISCUSSIONS

Cross-Lingual Transfer Performance

Table 2 presents the macro-F1 of both models across the transfer scenarios on the test sets of 600 comments per language.

Table 2. Macro-F1 across transfer scenarios (%)

Scenario	IndoBE RT BT	IndoBE RT MD	XLM-R BT	XLM-R MD
S1 zero-shot	48.2	61.5	51.7	63.0
S2 few-shot (100)	57.3	67.0	59.8	68.2
S2 few-shot (500)	66.4	72.8	68.9	73.5
S3 full	74.6	80.3	76.1	80.9

Three patterns emerge. First, zero-shot transfer from Indonesian worked much better for Melayu Deli than for Batak Toba, with a gap of more than 11 points for both models. This is consistent with linguistic distance: Melayu Deli is a Malay variety that shares much of its vocabulary and grammar with Indonesian, whereas Batak Toba belongs to a different branch of the Austronesian family with distinct vocabulary and morphology. The finding agrees with Lauscher et al. (2020), who reported that zero-shot transfer degrades sharply for typologically distant languages. Second, even a small number of target examples helped considerably; 500 examples recovered most of the gap between zero-shot and full fine-tuning for Melayu Deli. Third, XLM-R slightly outperformed IndoBERT in every setting, with the largest advantage on Batak Toba. A plausible explanation is that its multilingual vocabulary and exposure to several Austronesian languages give it subword representations that generalise better to unseen languages, whereas IndoBERT's representations are more specialised to Indonesian.

Effect of Data Augmentation

Table 3 compares the augmentation strategies using XLM-R in the full fine-tuning setting.

Table 3. Effect of augmentation (XLM-R, macro-F1, %)

Method	BT	MD
No augmentation (S3)	76.1	80.9
EDA	76.8	81.2
Back-translation	77.0	81.0
LLM, unfiltered	75.3	81.6
LLM, filtered (proposed)	79.4	83.1

EDA and back-translation produced only small gains of under one point, as their

variations remain close to the original sentences. Unfiltered LLM data gave mixed results: a modest improvement for Melayu Deli but a decrease of 0.8 points for Batak Toba. Filtered LLM data produced the best results in both languages, improving macro-F1 by 3.3 points for Batak Toba and 2.2 points for Melayu Deli over no augmentation, and the improvement was statistically significant ($p < 0.05$). At the class level, the filtered model achieved F1-scores of 77.8% (HS), 74.2% (ABS), and 86.2% (NT) for Batak Toba, and 82.5%, 78.9%, and 87.9% for Melayu Deli; the abusive class remained the hardest, as it often overlaps with hate speech when insults contain implicit identity references.

Quality of Synthetic Data

The difference between filtered and unfiltered data is explained by the quality analysis in Table 4.

Table 4. Quality of LLM-generated data

Indicator	BT	MD
Candidates generated	3,000	3,000
Retained after filtering	1,830 (61.0%)	2,220 (74.0%)
Fluency, audited (1–5)	3.1	4.0
Label fidelity, audited	78.0%	88.0%
Distinct-2 (synthetic / real)	0.41 / 0.58	0.47 / 0.55

The LLM wrote noticeably better Melayu Deli than Batak Toba. Native speakers rated Batak Toba outputs as only moderately fluent, often noting Indonesian words inserted into Batak sentence structures or unnatural word choices, and about one in five Batak Toba candidates had a label that did not match its content. Fewer Batak Toba candidates therefore survived filtering. The synthetic data were also less lexically diverse than real comments, as shown by the lower distinct-2 ratio, which suggests that LLM output tends to repeat certain phrasings. These observations explain why unfiltered augmentation hurt Batak Toba: the classifier was trained on a substantial share of disfluent or mislabelled examples. The result supports the concern in the literature that LLMs are unreliable in under-represented languages (Aji et al., 2022) and shows that quality control is not optional.

Bias Towards Identity Terms

Table 5 reports the identity-term false positive rate on neutral comments mentioning ethnic or religious groups.

Table 5. Identity-term false positive rate (% , lower is better)

Training data	BT	MD
No augmentation (S3)	18.4	14.0
LLM, unfiltered	24.1	19.3
LLM, filtered (proposed)	15.2	11.3

All models flagged a considerable share of harmless comments simply because they mentioned a group, reflecting the known tendency of hate speech classifiers to associate identity terms with toxicity (Sap et al., 2019). Unfiltered LLM data made this worse, increasing the false positive rate by 5.7 points for Batak Toba. Inspection showed that the LLM tended to produce hate speech examples that relied on a narrow set of identity terms, reinforcing a shortcut in which the mere mention of a group signals hate. After filtering, and because the generation prompts also requested neutral comments mentioning groups, the false positive rate fell below that of the original model. This is an important practical result, because a moderation tool that disproportionately flags discussions about particular ethnic or religious communities can itself become a source of unfair treatment.

Discussion and Limitations

Overall, the results suggest a practical recipe for low-resource hate speech detection in Indonesian regional languages: start from a multilingual model, collect even a few hundred native-speaker-labelled examples, and use LLM-generated data only with explicit filtering and a native-speaker audit. Several limitations remain. The corpus covers a limited time span and set of platforms, and hate speech is highly sensitive to current events. The two languages studied are relatively well-represented compared with many others in Indonesia, so results for more distant or less-documented languages may differ. The bias analysis covered ethnic and religious identity only, and a larger and more systematically constructed test set would allow stronger conclusions. Finally, LLM quality changes quickly, so the specific numbers in Table 4

should be re-measured whenever a different model is used.

CONCLUSION

This study introduced the first annotated hate speech corpus for Batak Toba and Melayu Deli and evaluated cross-lingual transfer and LLM-based augmentation for this task. Zero-shot transfer from Indonesian was effective for the closely related Melayu Deli but weak for Batak Toba, and XLM-R consistently outperformed IndoBERT. Filtered LLM augmentation achieved the best macro-F1, 79.4% for Batak Toba and 83.1% for Melayu Deli, and reduced false positives on neutral identity-related comments, whereas unfiltered synthetic data harmed Batak Toba performance and amplified bias. These findings indicate that LLMs can expand training data for low-resource languages only when combined with quality control and native-speaker involvement. Future work will extend the corpus to other North Sumatran languages such as Karo and Mandailing, explore parameter-efficient adaptation of LLMs directly as classifiers, and release the corpus and annotation guidelines to support further research.

BIBLIOGRAPHY

- Aji, A. F., Winata, G. I., Koto, F., Cahyawijaya, S., Romadhony, A., Mahendra, R., Kurniawan, K., Moeljadi, D., Prasajo, R. E., Baldwin, T., Lau, J. H., & Ruder, S. (2022). One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 7226–7249.
- Alfina, I., Mulia, R., Fanany, M. I., & Ekanata, Y. (2017). Hate speech detection in the Indonesian language: A dataset and preliminary study. *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 233–238.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners.

-
- Advances in Neural Information Processing Systems*, 33, 1877–1901.
- Cahyawijaya, S., Lovenia, H., Aji, A. F., Winata, G. I., Wilie, B., et al. (2023). NusaCrowd: Open source initiative for Indonesian NLP resources. *Findings of the Association for Computational Linguistics: ACL 2023*, 13745–13818.
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 8440–8451.
- Davidson, T., Warmusley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 512–515.
- Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., & Hovy, E. (2021). A survey of data augmentation approaches for NLP. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 968–988.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378–382.
- Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51(4), 1–30.
- Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., & Kamar, E. (2022). ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 3309–3326.
- Ibrohim, M. O., & Budi, I. (2019). Multi-label hate speech and abusive language detection in Indonesian Twitter. *Proceedings of the Third Workshop on Abusive Language Online*, 46–57.
- Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, 757–770.
- Lauscher, A., Ravishankar, V., Vulić, I., & Glavaš, G. (2020). From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4483–4499.
- Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. *International Conference on Learning Representations (ICLR)*.
- McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2), 153–157.
- Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual BERT? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 4996–5001.
- Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 1668–1678.
- Sennrich, R., Haddow, B., & Birch, A. (2016). Improving neural machine translation models with monolingual data. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 86–96.
- Wei, J., & Zou, K. (2019). EDA: Easy data augmentation techniques for boosting performance on text classification tasks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6382–6388.
- Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., & Purwarianti, A. (2020). IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding. *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the ACL and the 10th International Joint*

-
- | | |
|--|--|
| <p><i>Conference on NLP (ACL-IJCNLP)</i>, 843–857.</p> <p>Winata, G. I., Aji, A. F., Cahyawijaya, S., Mahendra, R., Koto, F., Romadhony, A., et al. (2023). NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages. <i>Proceedings of the 17th</i></p> | <p><i>Conference of the European Chapter of the Association for Computational Linguistics (EACL)</i>, 815–834.</p> <p>Add 5–10 recent (2021–2026) references from SINTA/Scopus-indexed journals on Indonesian hate speech detection and regional-language NLP.</p> |
|--|--|