Available online at :
http://ejournal.amikompurwokerto.ac.id/index.php/telematika/

Telematika

Accreditated SINTA "2" Kemenristek/BRIN, No. 85/M/KPT/2020

UNIVERSITAS AMIKOM PURWOKERTO

Fairness Auditing and Bias Mitigation in Aspect-Based Sentiment Models for Indonesian Public Services

Muhammad Shihab Fathurrahman Jondien[1,*], Taqwa Hariguna[2], Dhanar Intan Surya Saputra[3]

[1.2,3] Magister of Computer Science, Amikom Purwokerto University, Indonesia

ARTICLE INFO

History of the article:
Received August 20, 2025
Revised September 25, 2025
Accepted February 28, 2026

Keywords:
Fairness Auditing
Bias Mitigation
Aspect-Based Sentiment Analysis
Indobert,
Indonesian NLP
Ethical AI
Low-Resource Languages

Correspondence:

E-mail: fathurshihab@gmail.com

ABSTRACT

This study presents a comprehensive fairness audit and bias mitigation framework for Indonesian sentiment analysis using the SmSA IndoNLU dataset and the IndoBERT language model. The research investigates demographic and linguistic fairness by evaluating model performance across gender and regional groups and introduces an aspect-based extension to assess semantic fairness using an ABSA-style input formulation. Fairness metrics such as ΔF1, Demographic Parity Difference (DPD), and Equality of Opportunity were employed to quantify disparities in model behavior. The baseline IndoBERT model achieved strong overall accuracy (0.942) and macro-F1 (0.927) but exhibited significant regional bias, particularly toward Eastern and Sumatran dialects. A re-weighting strategy effectively reduced the regional F1 disparity by 59 percent with minimal accuracy loss, demonstrating the viability of loss-based fairness mitigation. The ABSA-style IndoBERT further improved fairness consistency across dialectal and aspect categories, achieving a macro-F1 of 0.930. Despite these improvements, aspect-level imbalances persisted, indicating that fairness challenges extend beyond demographic representation to semantic coverage. This work contributes an empirical and methodological foundation for ethical NLP evaluation in Bahasa Indonesia, emphasizing fairness auditing, bias mitigation, and responsible deployment of language models in low-resource and linguistically diverse settings.

1. INTRODUCTION

Natural Language Processing (NLP) has become increasingly important in digital communication and public service systems in Indonesia. Sentiment analysis models are widely applied in governance monitoring, complaint management, and social media analysis. The emergence of pretrained Indonesian models such as IndoBERT has accelerated research and deployment across various domains (Wilie et al., 2020; Koto et al., 2020). However, as these systems are integrated into socially sensitive contexts, concerns regarding fairness, bias, and ethical deployment have intensified (Venugopal et al., 2024; Fauzan & Saptawijaya, 2023).

Fairness in NLP refers to equitable model performance across demographic and linguistic groups. In a linguistically diverse country such as Indonesia, which has more than 700 local languages and dialects (Aji et al., 2022), unequal performance may marginalize certain communities and distort the interpretation

of public feedback. Although benchmarks such as IndoNLU and various fine-tuning approaches have improved overall performance (Ahmadian et al., 2024; Tandi et al., 2025), fairness analysis in Indonesian NLP remains limited. Existing corpora are often dominated by formal or urban language varieties, leading to disparities when models encounter underrepresented dialects or speech communities (Wongso et al., 2024; Purnomo & Sutopo, 2024). Most prior research emphasizes aggregate performance rather than disaggregated evaluation across demographic, regional, and semantic dimensions.

In addition, fairness studies in Indonesian NLP have primarily focused on demographic attributes, while aspect-level or semantic fairness has received less attention. Evaluating performance across thematic domains such as service, price, and facilities is particularly relevant in public service sentiment analysis, where topic imbalance may affect interpretability and accountability (Perwira et al., 2025; Febrianto et al., 2024).

This study conducts a structured fairness audit of IndoBERT-based sentiment models using the SmSA IndoNLU dataset. Model performance is evaluated across gender, regional, and aspect dimensions using F1 disparity, Demographic Parity Difference, and Equality of Opportunity. To address imbalance, a re weighting strategy is applied (Venugopal et al., 2024), and an Aspect Based Sentiment Analysis formulation is introduced to enable fine-grained semantic fairness evaluation (Wafda et al., 2025).

The contribution lies in providing a multidimensional fairness audit framework for Indonesian sentiment analysis and demonstrating that fairness improvements can be achieved with minimal impact on predictive performance. Integrating fairness auditing and mitigation into model evaluation supports the development of more transparent and inclusive NLP systems for Bahasa Indonesia.

2. LITERATURE REVIEW

Fairness in NLP is commonly divided into representational and allocative dimensions. Representational fairness concerns equitable model performance across social or linguistic groups, while allocative fairness refers to unequal downstream consequences of model outputs. Although most NLP studies focus on representational disparities due to measurability, such bias may translate into allocative impact in applied contexts such as public service monitoring. This distinction highlights the importance of structured fairness auditing in socially sensitive deployments.

Various fairness auditing and mitigation techniques have been proposed. Counterfactual evaluation and causal mediation approaches assess prediction stability when identity-related terms are modified (Da et al., 2024). Re weighting strategies address imbalance during training and are considered suitable for low resource settings (Tandi et al., 2025). Other approaches such as adversarial debiasing and data augmentation have shown improvements in sentiment and emotion classification (Christian et al., 2025), but they often require larger datasets and computational resources.

Fairness research has largely focused on high resource languages. In contrast, multilingual and low resource environments such as Indonesian face additional challenges due to dialectal diversity and uneven data representation. Benchmarks including IndoNLG, IndoNLI, and NusaBERT expand linguistic coverage (Cahyawijaya et al., 2021; Mahendra et al., 2021; Wongso et al., 2024), yet systematic fairness auditing

remains limited. Domain specific resources such as IndoGovBERT and IndoPref prioritize performance and coverage rather than disaggregated fairness evaluation (Riyadi et al., 2024; Wiyono et al., 2025).

Prior studies also indicate that IndoBERT based systems are sensitive to linguistic and topical variation. Domain specific fine tuning may reinforce topical imbalance (Syazali and Yulianti, 2025), and performance disparities have been observed in political and misinformation classification tasks under heterogeneous data distributions (Fathin et al., 2024; Praha et al., 2024). However, most Indonesian NLP research emphasizes aggregate accuracy rather than fairness across demographic, regional, or semantic subgroups (Ahmadian et al., 2024; Dwitama et al., 2023).

Overall, fairness research in Indonesian NLP remains underdeveloped, particularly in integrating demographic, regional, and aspect level analysis within a unified framework. Existing mitigation strategies provide a foundation, but systematic evaluation in linguistically diverse settings is still needed. This gap motivates a multidimensional fairness audit that combines bias measurement and mitigation within IndoBERT based sentiment modeling.

3. RESEARCH METHODS

This study proposes a structured experimental framework to audit and mitigate fairness bias in Indonesian sentiment analysis models. The methodology integrates dataset enrichment, demographic proxy labeling, fairness quantification, and bias mitigation strategies within a unified evaluation pipeline. IndoBERT was selected as the base model due to its broad coverage of Bahasa Indonesia and strong performance on IndoNLU benchmarks. Fairness was evaluated across gender, regional, and aspect-based dimensions to capture both social and semantic disparities.

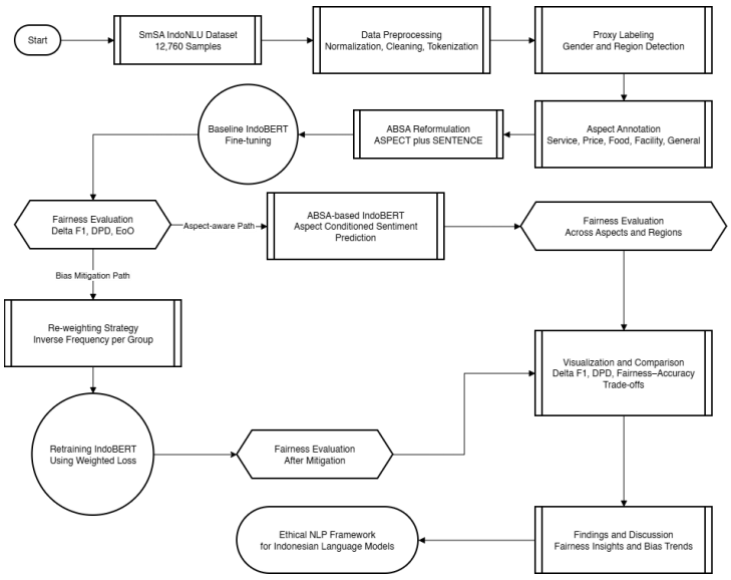

Figure 1. Research Flow

3.1. Dataset and Preprocessing

The experiments used the SmSA IndoNLU dataset, containing 12,760 labeled sentences in Bahasa Indonesia. Each sample was annotated as positive, neutral, or negative. The dataset mainly consists of public opinion and social media posts, making it suitable for sentiment analysis in public-service contexts.

Standard preprocessing steps were applied, including text normalization, lowercasing, and the removal of punctuation and URLs. The dataset was stratified into 80% training and 20% test partitions while maintaining sentiment balance across splits to ensure consistent evaluation conditions.

3.2. Proxy Demographic and Regional Attributes

Because the SmSA dataset lacks explicit demographic information, proxy attributes were generated based on lexical indicators. Gender was inferred using gendered expressions such as cewek (girl), perempuan (woman), cowok (boy), and laki-laki (man). Regional identity was estimated from dialectal cues, including arek and kowe (Javanese), urang and mah (Sundanese), ado and awak (Sumbanese), and beta or dong (Eastern Indonesian). Each text was automatically assigned a proxy label corresponding to the detected region or gender. It is important to emphasize that these proxy labels do not represent verified demographic identity. They serve as heuristic indicators to enable quantitative fairness analysis in the absence of human annotated demographic data. Consequently, the fairness results should be interpreted as performance disparities across linguistic signals rather than definitive demographic attributes. Although these heuristics are not perfectly representative of demographic identity, they enable a quantitative assessment of fairness disparities in the absence of annotated demographic data.

3.3. Aspect Annotation and ABSA Formulation

To incorporate semantic fairness evaluation, the dataset was augmented with synthetic aspect labels derived from rule-based keyword matching. Each text was categorized into one of five domains: service (pelayanan), price (harga), food (makanan), facility (fasilitas), or general (umum). The dataset was reformulated for Aspect-Based Sentiment Analysis (ABSA), where each input was structured as [ASPECT] [SEP] [SENTENCE]. This formulation allowed the model to capture aspect-specific sentiment polarity, facilitating fairness evaluation not only across demographics but also across semantic topics. The ABSA configuration enables fairness evaluation across thematic domains and provides additional interpretability by isolating aspect specific performance. The ABSA configuration enhances interpretability and reveals whether fairness disparities persist across different thematic aspects of public discourse.

3.4. Model Architecture and Fine-Tuning

The model used in this study was IndoBERT-base-p1, a transformer model pretrained on large-scale Indonesian text corpora. It contains twelve attention layers and 110 million parameters. The model was fine-tuned for three-way sentiment classification using a cross-entropy loss function. The training setup included a learning rate of 3×10^{-5}, batch size of 16, weight decay of 0.01, and four training epochs. Optimization was performed using the AdamW optimizer, and model selection was based on the best macro-F1 score on the validation set. To ensure comparability, identical hyperparameters were applied to the baseline, re-weighted, and ABSA-based models.

3.5. Fairness Metrics

Fairness was evaluated using three complementary metrics designed to capture both performance disparity and outcome imbalance.

F1 Disparity (ΔF1): The F1 disparity measures the range of F1 scores across demographic groups and is defined as

$$\Delta F1 = \max_{g \in G} (F1_g) - \min_{g \in G} (F1_g) \quad (1)$$

$F1_g$ represents the F1 score for group g, and G denotes the set of demographic or regional groups. A smaller ΔF1 indicates more equitable performance across groups.

Demographic Parity Difference (DPD): DPD quantifies differences in the probability of receiving a positive prediction across groups:

$$DPD = \max_{g \in G} P(\hat{Y} = 1 \mid A = g) - \min_{g \in G} P(\hat{Y} = 1 \mid A = g) \quad (2)$$

$\hat{Y}$ is the predicted label and A is the sensitive attribute (e.g., gender or region). Lower DPD values correspond to more balanced output distributions.

Equality of Opportunity (EoO): EoO evaluates the fairness of true positive rates between groups and is computed as

$$EoO = \max_{g \in G} P(\hat{Y} = 1 \mid Y = 1, A = g) - \min_{g \in G} P(\hat{Y} = 1 \mid Y = 1, A = g) \quad (3)$$

Y is the ground-truth label. A low EoO difference implies equal sensitivity across demographic subgroups.

Together, these metrics provide a multidimensional assessment of fairness across predictive performance and outcome distribution.

3.6. Bias Mitigation through Re-weighting

To reduce the effect of demographic imbalance, a re-weighting strategy was implemented during training. Each instance in the training data was assigned a weight inversely proportional to the frequency of its group:

$$w_i = \frac{1/p_{A_i}}{\frac{1}{N}\sum_{j=1}^{N}(1/p_{A_j})} \quad (4)$$

w_i denotes the sample weight for instance i, p_{A_i} is the empirical probability of the group A_i, and N is the total number of training samples. This normalization ensures that the average weight equals one, preserving training stability. The loss function was modified to incorporate these sample weights:

$$L = \frac{1}{N}\sum_{i=1}^{N} w_i \cdot \text{CrossEntropy}(f_\theta(x_i), y_i) \quad (5)$$

$f_\theta(x_i)$ represents the model prediction for input x_i and y_i is the true sentiment label. This approach amplifies the contribution of minority samples while maintaining balanced optimization dynamics, effectively mitigating representational bias in the training process.

3.7. Evaluation Framework

The evaluation consisted of three configurations: the baseline IndoBERT, the re-weighted IndoBERT, and the ABSA-based IndoBERT. Each configuration was trained under identical settings to ensure comparability. Model performance was evaluated using accuracy and macro-F1, while fairness was assessed through ΔF1, DPD, and EoO for gender, region, and aspect attributes. Visualization analyses were generated to examine fairness–performance trade-offs, and counterfactual testing was conducted by swapping gender-related terms (for example, cewek and cowok) to assess prediction stability. This

comprehensive evaluation protocol enabled both quantitative and qualitative analysis of fairness behavior in Indonesian sentiment models, providing insights into the trade-offs between accuracy and equity.

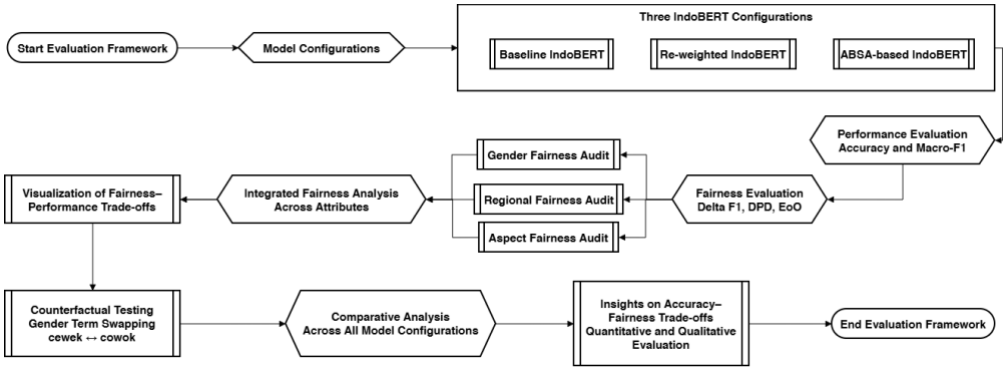

Figure 2. Evaluation Framework

4. RESULTS AND DISCUSSION

4.1. Dataset and Attribute Distribution

The experiments used the SmSA IndoNLU dataset, which contains 12,760 Indonesian public opinion texts labeled as positive, neutral, or negative. To enable fairness analysis, the dataset was extended with synthetic aspect labels and proxy demographic attributes for gender and region.

Aspect labels were assigned using keyword-based rules reflecting common public service themes: Service (*pelayanan*), Price (*harga*), Food (*makanan*), Facility (*fasilitas*), and General (*umum*). Sentences containing relevant keywords such as *pelayanan* (service), *pegawai* (employee), or *PNS* (civil servant) were categorized as Service. Mentions of *harga* (price), *mahal* (expensive), or *murah* (cheap) were assigned to Price. References to *makanan* (food), *resto* (restaurant), or *warung* (food stall) were labeled as Food. Texts containing *fasilitas* (facility), *toilet* (toilet), or *parkir* (parking) were grouped under Facility. All remaining texts were categorized as General.

Demographic proxies were inferred from lexical cues. Gender indicators included terms such as *cewek* (girl) and *perempuan* (woman) for female, and *cowok* (boy) and *laki-laki* (man) for male. Regional identity was approximated using dialectal markers associated with Sundanese, Javanese, Sumatran, and Eastern Indonesian speech patterns, such as *urang* and *mah* (Sundanese), *arek* and *kowe* (Javanese), *lah* and *ado* (Sumatran), and *beta* or *dong* (Eastern Indonesian).

Table 1 shows substantial imbalance across attributes. Nearly half of the dataset (49.6 percent) belongs to the General aspect, whereas Facility and Service represent only 6.4 percent and 8.4 percent of the data. Gender labels are highly skewed, with 97.6 percent categorized as unknown and fewer than 2.5 percent identified as male or female. Regional distribution is also uneven, with Eastern dialectal markers dominating at 45.3 percent, while Javanese markers appear in only 0.3 percent of samples.

Table 1. Dataset and Attribute Distribution

Attribute	Category	Count	% of Total
Aspect	General	6,334	49.6%
	Price	2,280	17.9%

	Food	2,258	17.7%
	Service	1,068	8.4%
	Facility	820	6.4%
Gender	Male	197	1.5%
	Female	111	0.9%
	Unknown	12,452	97.6%
Region	Eastern	5,788	45.3%
	Sundanese	2,434	19.1%
	Sumatran	2,333	18.3%
	Javanese	37	0.3%
	Unknown	2,168	17.0%

These distributions indicate strong representational imbalance across demographic, regional, and thematic dimensions, underscoring the importance of fairness auditing in Indonesian NLP applications.

4.2. Baseline IndoBERT Fairness Audit

To establish a baseline for fairness evaluation, we fine-tuned the IndoBERT-base-p1 model on the SmSA dataset for sentiment classification. The fine-tuned model achieved a Macro-F1 score of 0.927 and an overall accuracy of 0.942, indicating robust performance in terms of aggregate predictive quality. However, when the model's predictions were disaggregated by demographic attributes, significant disparities emerged across gender and regional subgroups. As presented in Table 2, IndoBERT's performance varied notably between linguistic groups. The model performed consistently well on samples associated with the Sundanese dialect (F1 = 0.891) and Javanese expressions (F1 = 0.925), but its performance dropped sharply on texts containing Eastern Indonesian linguistic markers (F1 = 0.805). This decline suggests that the model's representation space may be biased toward more prevalent linguistic patterns in the IndoNLU training corpus, leading to weaker generalization for underrepresented dialects. In terms of gender, the difference in F1 scores between male and female texts appears minimal, with both achieving near-perfect performance. However, given the extremely small number of female-identified samples (n = 14), such results should be interpreted with caution, as statistical reliability is limited by data scarcity.

Table 2. Baseline Fairness Results by Gender and Region

Attribute	Group	F1	Accuracy	Positive Rate	Count
Gender	Male	0.916	0.969	0.718	458
	Female	1.000	1.000	0.857	14
Region	Javanese	0.925	0.955	0.522	67
	Sundanese	0.891	0.932	0.841	44
	Eastern	0.805	0.913	0.130	23

The fairness metrics derived from these results reinforce the presence of linguistic bias in the model. The observed regional disparity (ΔF1 = 0.121) indicates that IndoBERT's ability to detect sentiment differs significantly across regional linguistic variants. Similarly, the Demographic Parity Difference (DPD = 0.710) suggests a substantial imbalance in positive sentiment predictions between regions, with Eastern dialects receiving markedly fewer positive classifications. This trend aligns with known limitations in pretrained Indonesian models, which are typically trained on data dominated by urban or Western Indonesian language varieties. The visualized results in Figure 3 further highlight the magnitude of this bias, showing that regional disparities constitute the most prominent dimension of fairness imbalance,

whereas gender-based bias remains comparatively minor. These findings emphasize the need for fairness mitigation strategies that target dialectal and regional diversity within Indonesian NLP models.

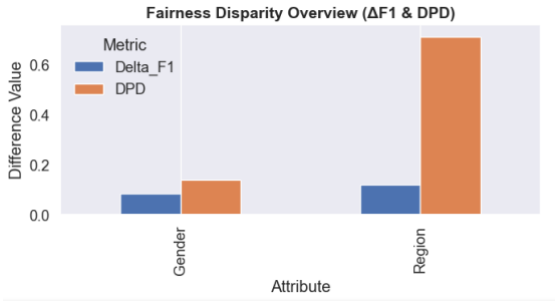

Figure 3. Fairness Disparity Overview (ΔF1 and DPD)

Regional bias dominates the fairness disparities across demographic dimensions. The bar chart compares the difference in F1 score (ΔF1) and Demographic Parity Difference (DPD) for gender and regional attributes.

4.3. Bias Mitigation via Re-weighting

To address representational imbalance identified in the baseline audit, a re-weighting strategy was applied during training. Each training instance was assigned a weight inversely proportional to the frequency of its demographic group. This adjustment increased the contribution of underrepresented categories, particularly female-associated texts and Eastern Indonesian dialectal expressions, without altering the dataset composition.

As shown in Table 3, re-weighting substantially reduced regional disparity. The regional F1 gap (ΔF1) decreased from 0.121 to 0.049, corresponding to an improvement of approximately 59 percent. Macro-F1 remained stable, with only a slight decrease from 0.927 to 0.924, indicating minimal impact on overall predictive performance.

Table 3. Fairness Comparison Before and After Re-weighting

Metric	Gender (Before → After)	Region (Before → After)	Δ (Improvement)
ΔF1	0.084 → 0.128	0.121 → 0.049	Improved (Region)
DPD	0.139 → 0.212	0.710 → 0.777	Slight increase
Macro-F1	0.927 → 0.924	–	Stable

However, Demographic Parity Difference increased slightly for both gender and region, suggesting changes in the distribution of positive predictions across groups. This reflects a trade-off between performance parity and distributional parity, where reducing disparity in F1 scores may affect output calibration. The increase in gender ΔF1 from 0.084 to 0.128 is likely influenced by the very small number of female-identified samples, which limits reliable subgroup generalization. Overall, re-weighting effectively reduced regional bias while maintaining predictive stability, demonstrating its practicality as a mitigation strategy for fairness enhancement in low-resource and linguistically diverse settings.

Figure 4 further illustrates that regional F1 disparities decreased after mitigation, while DPD increased marginally. These findings highlight the importance of evaluating multiple fairness criteria when applying mitigation strategies in multilingual and low-resource NLP contexts.

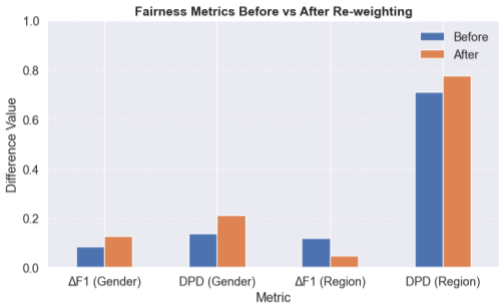

Figure 4. Fairness Metrics Before and After Re-weighting

4.4. ABSA-style IndoBERT: Aspect-based Fairness

To evaluate fairness across semantic dimensions, IndoBERT was extended using an Aspect-Based Sentiment Analysis formulation. Each input was structured as:[ASPECT] [SEP] [SENTENCE]. This design enables the model to condition sentiment prediction on a specific thematic aspect. The ABSA-based model achieved an overall accuracy of 0.946 and a macro-F1 of 0.930, indicating strong contextual performance across public service themes.

However, aspect-level evaluation (Table 4) reveals noticeable variation across categories. The highest F1 scores were observed for Service (0.941) and General (0.932), which have relatively larger sample sizes. In contrast, lower performance was found for Facility (0.785) and Food (0.819), resulting in an aspect-level disparity of approximately 0.16. This pattern suggests that underrepresented thematic categories receive weaker representation during training, leading to semantic imbalance. Although overall accuracy remains high, the results indicate that data distribution directly influences fairness across topics.

Table 4. ABSA Model Performance by Aspect

Aspect	F1	Accuracy	Count
Service	0.941	0.962	208
General	0.932	0.939	1,274
Food	0.819	0.961	439
Price	0.873	0.946	464
Facility	0.785	0.946	167

Aspect-level F1 scores show that minor aspects such as Facility and Food suffer from weaker representation and lower performance. The figure presents a bar chart comparing F1 scores for each aspect category.

4.5. Fairness Across Gender and Region (ABSA-style)

The ABSA formulation contributes not only to interpretability but also to improved fairness stability across demographic and regional groups. By conditioning predictions on aspect tokens, the model demonstrates more consistent performance across subgroups.

As shown in Table 5, gender-related metrics reached perfect scores for both male and female categories. However, these results must be interpreted cautiously due to the very small number of gender-identified samples. While no observable disparity appears in this configuration, statistical reliability remains limited.

More substantial improvements were observed across regional groups. The ABSA-based model achieved F1 scores of 0.917 for Eastern, 0.886 for Sumatran, and 0.918 for Sundanese categories, reducing the regional performance gap (ΔF1) to 0.032. This represents a significant improvement compared to both the baseline and re-weighted models.

The findings suggest that aspect-level conditioning reduces reliance on dialectal or surface lexical cues and promotes more stable sentiment representations across sociolinguistic variations. The fairness and performance comparison further indicates that the ABSA formulation maintains competitive macro-F1 while achieving lower regional disparity, demonstrating a favorable balance between predictive accuracy and equity.

Table 5. ABSA Fairness by Gender and Region

Attribute	Group	F1	Accuracy	Count
Gender	Male	1.000	1.000	47
	Female	1.000	1.000	19
Region	Eastern	0.917	0.943	1,176
	Sumatran	0.886	0.943	471
	Sundanese	0.918	0.952	475

The comparison across baseline, re-weighted, and ABSA configurations shows that incorporating semantic context through aspect tokens reduces regional disparity while preserving overall performance.

4.6. Summary of Fairness Across Dimensions

Table 6 summarizes fairness outcomes across gender, regional, and aspect dimensions. Regional disparity showed the most substantial improvement following re-weighting and ABSA conditioning, with ΔF1 reduced to 0.049. This indicates more consistent performance across dialectal groups. However, Demographic Parity Difference for region remained relatively high (0.777), suggesting persistent imbalance in the distribution of positive predictions despite improved predictive consistency.

Table 6. Summary of Fairness Metrics Across Dimensions

Fairness Dimension	ΔF1	DPD	Observation
Gender	0.128	0.212	Stable but underrepresented
Region	0.049	0.777	Improved fairness, DPD remains high
Aspect	0.156	–	Persistent aspect imbalance bias

Gender-related metrics appeared stable (ΔF1 = 0.128, DPD = 0.212), but interpretation remains limited due to the very small number of gender-identified samples. As a result, reliable conclusions regarding gender fairness cannot be firmly established.

The most persistent disparity was observed in the aspect dimension (ΔF1 = 0.156). The model consistently performed better on more frequent aspects such as General and Service compared to underrepresented categories such as Facility and Food. This finding indicates that fairness challenges extend beyond demographic and regional factors to include semantic imbalance within the dataset.

Overall, mitigation strategies were most effective for regional fairness, while aspect-level disparity remains an open challenge. These results highlight the need to address both representational and semantic imbalance in Indonesian NLP systems.

4.7. Discussion

The results indicate that regional disparity represents the most prominent fairness imbalance in Indonesian sentiment analysis models. The baseline IndoBERT exhibited lower performance for Eastern and Sumatran dialectal expressions compared to more dominant varieties, reflecting patterns previously observed in Indonesian pretrained models (Purnomo & Sutopo, 2024; Wongso et al., 2024). In a linguistically diverse context such as Indonesia (Aji et al., 2022), limited dialectal representation in training data can lead to uneven generalization, constituting representational bias.

Re-weighting substantially reduced regional F1 disparity while maintaining overall predictive performance, consistent with prior findings that training-level adjustments can mitigate imbalance without major accuracy loss (Venugopal et al., 2024; Tandi et al., 2025). However, the increase in Demographic Parity Difference after mitigation demonstrates that performance parity does not necessarily ensure distributional parity, echoing broader fairness discussions in sentiment modeling (Da et al., 2024). These results emphasize the need to evaluate multiple fairness metrics rather than relying on a single indicator.

The ABSA formulation further improved stability across regions by conditioning predictions on semantic aspects. Prior studies show that aspect-based modeling enhances contextual sensitivity in Indonesian sentiment tasks (Febrianto et al., 2024; Wafda et al., 2025), and domain-specific fine-tuning can influence representation learning in transformer models (Syazali & Yulianti, 2025). The reduced regional disparity under ABSA suggests that incorporating semantic structure may decrease reliance on dialect-correlated lexical cues.

Aspect-level imbalance remained a persistent challenge. Lower performance for Facility and Food categories compared to more frequent aspects reflects thematic underrepresentation, consistent with findings in other Indonesian classification tasks (Fathin et al., 2024; Praha et al., 2024). This indicates that fairness concerns extend beyond demographic and regional attributes to include semantic imbalance within datasets.

Gender-related findings require cautious interpretation. Although subgroup scores appeared similar, the extremely small number of female-identified samples limits statistical reliability. Moreover, demographic attributes were inferred through lexical proxies, meaning the evaluation reflects disparities across linguistic indicators rather than verified identities (Fauzan & Saptawijaya, 2023).

Overall, fairness improvements were achieved with minimal reduction in macro-F1, supporting prior evidence that fairness-oriented adjustments can be integrated without substantial efficiency loss (Christian et al., 2025; Venugopal et al., 2024). Nevertheless, fairness should not be reduced to performance metrics alone. In socially sensitive applications such as hate speech detection or public complaint monitoring (Dwitama et al., 2023; Yefferson et al., 2024), representational disparities may translate into allocative consequences.

These findings highlight the importance of incorporating fairness auditing into Indonesian NLP benchmarks. While resources such as IndoNLG and IndoNLI expand task coverage (Cahyawijaya et al., 2021; Mahendra et al., 2021), systematic fairness evaluation remains limited. Future work should strengthen statistical validation, demographic annotation practices, and cross-domain evaluation using emerging resources such as IndoPref and IndoGovBERT (Wiyono et al., 2025; Riyadi et al., 2024).

5. CONCLUSION

This study presented a comprehensive fairness audit and bias mitigation framework for Indonesian sentiment analysis models using the SmSA IndoNLU dataset. By extending IndoBERT with both demographic fairness evaluation and aspect-based contextualization, the research provides an empirical foundation for understanding how social, linguistic, and semantic imbalances influence model behavior in low-resource settings. The findings show that IndoBERT achieves strong overall performance in sentiment classification; however, substantial disparities persist across regional dialects and semantic aspects. The baseline analysis revealed that the model underperforms on Eastern and Sumatran linguistic expressions, reflecting the uneven distribution of dialectal data in Indonesian NLP corpora. Through the application of a re-weighting mitigation strategy, these regional disparities were significantly reduced without compromising overall predictive accuracy. Moreover, the ABSA-based IndoBERT formulation improved the consistency of sentiment prediction across regions by conditioning the model on explicit semantic contexts.

Despite these advances, the study also highlights persistent challenges in aspect-level and representational fairness. The aspect-based evaluation showed that underrepresented topics such as Facility and Food continue to exhibit lower predictive performance, illustrating that fairness issues can extend beyond demographic variables into semantic domains. Gender fairness remained inconclusive due to the limited availability of gender-identified data, underscoring the importance of more inclusive and balanced dataset design. The findings collectively demonstrate that fairness and accuracy are not mutually exclusive; equitable model performance can be achieved through targeted mitigation strategies that balance learning across linguistic and demographic dimensions.

In conclusion, this research contributes to the development of a methodological framework for ethical NLP evaluation in Bahasa Indonesia, emphasizing fairness metrics such as ΔF1, Demographic Parity Difference, and Equality of Opportunity. The study advocates for the integration of bias auditing, mitigation, and transparency reporting as standard practices in Indonesian NLP development. Future work should explore advanced mitigation techniques such as counterfactual data augmentation, adversarial debiasing, and post-hoc calibration to further enhance fairness consistency across demographic and semantic contexts. Establishing such practices will be critical to building trustworthy, inclusive, and socially responsible language technologies for multilingual and low-resource communities.

REFERENCES

Ahmadian, H., Abidin, T. F., Riza, H., & Muchtar, K. (2024). *Hybrid Models for Emotion Classification and Sentiment Analysis in Indonesian Language.* Applied Computational Intelligence and Soft Computing. https://doi.org/10.1155/2024/2826773

Aji, A. F., Winata, G. I., Koto, F., Cahyawijaya, S., et al. (2022). *One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia.* ArXiv. https://doi.org/10.48550/arxiv.2203.13357

Cahyawijaya, S., Winata, G. I., Wilie, B., et al. (2021). *IndoNLG: Benchmark and Resources for Evaluating Indonesian Natural Language Generation.* EMNLP. https://doi.org/10.18653/v1/2021.emnlp-main.699

Christian, W., Adamlu, D., Yu, A., & Suhartono, D. (2025). *Leveraging IndoBERT and DistilBERT for Indonesian Emotion Classification in E-Commerce Reviews.* arXiv. https://doi.org/10.48550/arxiv.2509.14611

Da, Y., Bossa, M. N., Berenguer, A. D., & Sahli, H. (2024). *Reducing Bias in Sentiment Analysis Models Through Causal Mediation Analysis and Targeted Counterfactual Training.* IEEE Access. https://doi.org/10.1109/access.2024.3353056

Dwitama, A. P. J., Fudholi, D. H., & Hidayat, S. (2023). *Indonesian Hate Speech Detection Using Bi-LSTM and IndoBERT.* Jurnal RESTI. https://doi.org/10.29207/resti.v7i2.4642

Fauzan, M. A., & Saptawijaya, A. (2023). *Analysis and Mitigation of Religion Bias in Indonesian NLP Datasets.* Jurnal RESTI. https://doi.org/10.29207/resti.v7i4.5035

Fathin, M. A., Sibaroni, Y., & Prasetyowati, S. (2024). *Handling Imbalance Dataset on Hoax Indonesian Political News Classification Using IndoBERT.* Jurnal Media Informatika Budidarma. https://doi.org/10.30865/mib.v8i1.7099

Febrianto, D., Fitriani, M. A., Afrad, M., & Khadija, M. A. (2024). *Aspect-Based Sentiment Analysis Menggunakan IndoBERT.* Melek IT Journal. https://doi.org/10.30742/melekitjournal.v10i2.358

Istiqomah, N., & Novika, F. (2025). *Comparative Performance of IndoBERT and IndoLEM for Post-Disaster Health Information Extraction.* Journal of Computer Science and Informatics Engineering. https://doi.org/10.55537/cosie.v4i3.1174

Khairunnisa, S. O., Chen, Z., & Komachi, M. (2023). *Dataset Enhancement and Multilingual Transfer for Named Entity Recognition in Indonesian Language.* ACM Transactions on Asian and Low-Resource Language Information Processing. https://doi.org/10.1145/3592854

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). *IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP.* COLING. https://doi.org/10.18653/v1/2020.coling-main.66

Mahendra, R., Aji, A. F., Louvan, S., et al. (2021). *IndoNLI: A Natural Language Inference Dataset for Indonesian.* EMNLP. https://doi.org/10.18653/v1/2021.emnlp-main.821

Perwira, R., Permadi, V. A., Purnamasari, D. I., & Agusdin, R. P. (2025). *Domain-Specific Fine-Tuning of IndoBERT for Aspect-Based Sentiment Analysis in Indonesian Travel UGC.* JISEBI. https://doi.org/10.20473/jisebi.11.1.30-40

Praha, T. C., Widodo, W., & Nugraheni, M. (2024). *Indonesian Fake News Classification Using Transfer Learning in CNN and LSTM.* JOIV: International Journal on Informatics Visualization.

Purnomo, T. D., & Sutopo, J. (2024). *Comparison of Pre-Trained BERT-Based Transformer Models for Regional Language Text Sentiment Analysis in Indonesia.* IJST.

Riyadi, A., Kovács, M., Serdült, U., & Kryssanov, V. (2024). *IndoGovBERT: A Domain-Specific Language Model for Processing Indonesian Government SDG Documents.* Big Data and Cognitive Computing. https://doi.org/10.3390/bdcc8110153

Syazali, M. R., & Yulianti, E. (2025). *Classification of Economic Activities in Indonesia Using IndoBERT Language Model.* Jurnal Ilmu Komputer dan Informasi. https://doi.org/10.21609/jiki.v18i2.1446

Tandi, T. Y., Abidin, T. F., & Riza, H. (2025). *Incorporation of IndoBERT and Machine Learning Features to Improve Indonesian RTE.* JISEBI. https://doi.org/10.20473/jisebi.11.2.173-186

Venugopal, J. P., Subramanian, A. A. V., Sundaram, G., Rivera, M., & Wheeler, P. W. (2024). *A Comprehensive Approach to Bias Mitigation for Sentiment Analysis of Social Media Data.* Applied Sciences. https://doi.org/10.3390/app142311471

Wafda, A., Fudholi, D., & Nugraha, J. (2025). *Aspect-Based Sentiment Analysis on Twitter Tweets about the Merdeka Curriculum Using IndoBERT.* JITK. https://doi.org/10.33480/jitk.v10i3.5692

Wilie, B., Vincentio, K., Winata, G. I., et al. (2020). *IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding.* AACL.

Wiyono, V. R., Anugraha, D., Purwarianti, A., & Winata, G. I. (2025). *IndoPref: A Multi-Domain Pairwise Preference Dataset for Indonesian.* arXiv. https://doi.org/10.48550/arxiv.2507.22159

Wongso, W., Setiawan, D. S., Limcorn, S., & Joyoadikusumo, A. (2024). *NusaBERT: Teaching IndoBERT to Be Multilingual and Multicultural.* arXiv. https://doi.org/10.48550/arxiv.2403.01817

Yefferson, D. Y., Lawijaya, V., & Girsang, A. S. (2024). *Hybrid Model: IndoBERT and Long Short-Term Memory for Detecting Indonesian Hoax News.* IAES International Journal of Artificial Intelligence.