Journal of Mathematics and Scientific Computing with Applications

E-ISSN : 2580-5754; P-ISSN : 2580-569X
Volume 7, Number 1, 2026
DOI: 10.53806/jmscowa.v7i1.1561
Page: 168–178

PANCASILA-BASED ASPECT CATEGORY SENTIMENT ANALYSIS FOR DETECTING NEGATIVE CONTENT IN INDONESIAN CODE-MIXED SOCIAL MEDIA

Stefani Tasya Hallatu[1], **Ratih Nur Esti Anggraini**[2], **Adhatus Solichah Ahmadiyah**[3]

[1,2,3]Informatics, Institut Teknologi Sepuluh Nopember Surabaya, Surabaya, 60111, Indonesia

Article Info

Article history:
Received : April 14, 2026
Revised : May 14, 2026
Accepted : July 30, 2026

Keywords:
Aspect Category Sentiment Analysis;
Code-Mixed;
Fine-Tuned Transformer;
Pancasila;
Zero-Shot Annotation.

ABSTRACT

Hate speech and value-violating content on Indonesian social media, compounded by code-mixed language, threaten social cohesion. This study proposes a Pancasila-based Aspect Category Sentiment Analysis framework grounded in Indonesia's five foundational values: Divinity, Humanity, Unity, Democracy, and Social Justice. Four Transformer models were evaluated under Full Fine-Tuning, LoRA, and QLoRA on 41,138 Indonesian code-mixed texts (confidence ≥ 0.75), annotated via zero-shot LLM inference and validated by two independent experts ($\kappa = 0.82$; LLM-expert $\kappa = 0.81$). IndoBERT LoRA achieved the highest in-pipeline F1-Score (0.77), though bootstrap intervals show this is statistically indistinguishable from several top configurations. Against an independent expert-validated ground truth (n = 1,000), all models dropped 6.7% on average; IndoBERTweet QLoRA obtained the highest point-estimate generalization (GT F1 = 0.71, 10.27 MB adapter storage), best read as the top of a statistically indistinguishable cluster rather than a confirmed single best model. The framework offers a culturally grounded approach to AI-assisted content moderation in Indonesia.

Corresponding Author:
Ratih Nur Esti Anggraini
Informatics
Institut Teknologi Sepuluh Nopember
Email: ratih_nea@if.its.ac.id

1. INTRODUCTION

Indonesia ranks among the world's largest social media markets, with 143 million active users [1]. This growth has enabled civic participation but has also fueled hate speech, abusive language, and discriminatory content [2], [3], prompting regulatory responses such as the Electronic Information and Transactions Law (UU ITE No. 19/2016) [4] and positioning AI-assisted moderation as a governance tool that must be interpretable and culturally accountable [3], [2]. A key complication is code-mixing between Indonesian, regional languages, and English [5], which

introduces syntactic irregularities and slang that conventional NLP models struggle to capture [6], compounded by the sheer volume of content that renders manual moderation infeasible at scale.

Transformers such as IndoBERT and IndoBERTweet perform well on Indonesian code-mixed ABSA/ACSA [7], fine-tuned mBERT improves ABSA and reduces out-of-vocabulary issues [8], cross-attention RoBERTa has improved ACSA [9], and AraBERT with text augmentation improves aspect detection in low-resource settings [10]. Parameter-efficient fine-tuning, LoRA [11] and QLoRA [12], offers computationally efficient alternatives to full fine-tuning with competitive performance in low-resource language scenarios [13].

Existing negative-content detection largely relies on generic taxonomies (binary hate speech, universal sentiment polarity) that ignore Indonesia's specific societal values: civil-toned content that undermines democratic deliberation or reinforces inequality can evade such taxonomies while still violating Indonesia's normative value system [14], [4]. Pancasila, Indonesia's constitutional state philosophy since 1945, offers a nationally legitimate, legally grounded alternative: Deity, Humanity, Unity, Democracy, and Social Justice [14]. Using these as an aspect taxonomy (i) enables fine-grained detection along culturally meaningful dimensions rather than coarse polarity, (ii) aligns classification output with Indonesia's regulatory discourse, and (iii) provides a reusable, institutionally grounded label set rather than an ad hoc one [14], [4]. To our knowledge, no prior study integrates ACSA with Pancasila values, compares multiple Indonesian and multilingual Transformers under identical conditions, or applies LoRA/QLoRA to Indonesian ACSA [13].

This study proposes a Pancasila-based ACSA framework for Indonesian code-mixed social media, fine-tuning four Transformers (IndoBERT, IndoBERTweet, mBERT, XLM-RoBERTa) under three strategies (FFT, LoRA, QLoRA; 12 configurations). Contributions: (1) a culturally grounded ACSA framework anchored in Pancasila as an alternative to generic hate-speech taxonomies; (2) a reproducible comparative evaluation of Indonesian-specific and multilingual Transformers across predictive performance and computational efficiency; and (3) a large-scale pseudo-labeled dataset validated through expert inter-annotator agreement analysis.

2. RESEARCH METHOD

This study used a quantitative computational design with seven stages: data collection, exploratory data analysis, preprocessing, zero-shot annotation, expert validation, model training, and comparative evaluation (Figure 1), enabling systematic comparison of architectures and fine-tuning strategies under identical conditions.

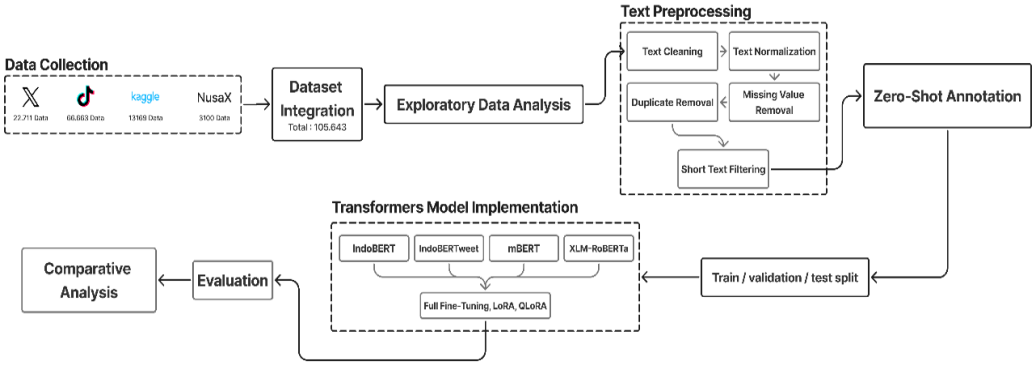

Figure 1. Research Methodology Flowchart

2.1. Data and Preprocessing

Data (105,643 raw instances) were collected from TikTok, X/Twitter (via APIFY, targeting politics, SARA, and gender/sexuality hashtags), the Indonesian Abusive and Hate Speech Kaggle dataset, and NusaX [15] (Table 1). All crawled text was anonymized; given the non-interventional, purely observational use of anonymized public text, formal IRB review was not required. NusaX is the only source of regional-language coverage (Javanese, Minangkabau, Batak) but forms just 6.98% of the preprocessed corpus, while TikTok remains dominant (∼62–63%) throughout preprocessing (duplicate-text rates: TikTok 27.8%, X 91.4%, vs. 1.1% Kaggle and 3.2% NusaX). Consequently "Indonesian code-mixed" here denotes colloquial Indonesian–English mixing typical of TikTok/X commentary, not general regional-language mixing, and generalization claims are bounded accordingly (revisited in Section 4). Cleaning removed URLs, mentions, hashtags, emojis, and non-alphabetic symbols; text was lowercased, normalized via a slang dictionary [6], deduplicated, and filtered to ≥5 tokens, leaving 42,223 preprocessed instances.

Table 1: Raw and Preprocessed Data Composition

Source	**Raw (n)**	**Preprocessed (%, $n = 42{,}223$)**
TikTok	66,663	62.08
X (Twitter)	22,711	4.16
Kaggle	13,169	26.77
NusaX	3,100	6.98
Total	**105,643**	**100.00**

2.2. Zero-Shot Annotation and Confidence Filtering

In the absence of a Pancasila-annotated benchmark, zero-shot annotation used Llama 3 8B via Ollama, selected given the 8 GB VRAM constraint of the local RTX 5060 GPU (which precludes larger models such as Llama 3.3 70B). Each text received one or more Pancasila aspect labels (multi-label; Table 3), operationalized per [14], and one sentiment label (Negative/Neutral/Positive), plus a confidence score. Before threshold filtering, 1,000 instances were stratified-sampled for expert validation (below), leaving a pool of 41,223 non-expert instances. A confidence threshold of ≥ 0.75 was chosen after comparing candidates: higher thresholds severely under-represented the Neutral class (down to 7.6% vs. 34.0% at ≥ 0.75), while ≥ 0.75 retained 41,138 of 41,223 instances (99.79%). Of the 85 excluded instances, 84 (98.8%) carried an explicit annotation-failure flag (confidence fixed at 0.0), i.e. exclusions were predominantly technical parsing failures rather than genuinely ambiguous cases, limiting selection-bias concerns [16], since prior work indicates pseudo-label quality matters more than quantity for downstream performance [16]. Table 2 details the discard pattern by source, text length, and sentiment: discards concentrate in TikTok (71/85) and Kaggle (14/85), with none from NusaX or X, and disproportionately in short texts (6–10 tokens: 59/85) and the Neutral class (84/85) – consistent with a technical-failure explanation rather than a systematic linguistic pattern.

Table 2: Discarded vs. Retained Instances by Source, Length, and Sentiment (Threshold ≥ 0.75, $n_{\text{pool}} = 41{,}223$)

Category	**Discarded**	**Retained**	**Rate (%)**
By Source			
TikTok	71	25,890	0.27
Kaggle	14	11,041	0.13
NusaX / X (Twitter)	0	4,207	0.00
By Text Length (tokens)			
6-10	59	16,197	0.36
11-20	19	14,646	0.13
>20	7	10,295	0.07
By Sentiment			
Negative / Positive	1	27,233	0.00
Neutral	84	13,905	0.60
Total	**85**	**41,138**	**0.21**

Table 3: Aspect Definitions and Confidence-Threshold Comparison ($n_{\text{pool}} = 41{,}223$)

Aspect	**Definition**
Deity	Violates religious tolerance; religion-based hate speech.
Humanity	Demeans human dignity; bullying, harassment, discrimination.
Unity	Racism, regional/ethnic hostility, threats to national unity.
Democracy	Undermines democratic/deliberative decision-making.
Social Justice	Reinforces social inequality or digital discrimination.

Threshold	≥ 0.95	≥ 0.85	≥ 0.75	≥ 0.65
Retained (%)	0.8	70.0	**99.79**	99.79
Neg/Net/Pos (%)	3/0/96	60/8/33	43/34/23	43/34/23

2.3. Expert Annotation Validation

A stratified sample of 1,000 instances (drawn before confidence filtering, no overlap with the 41,138-instance training pool) was independently labeled by two expert annotators (native Indonesian speakers with computational-linguistics backgrounds). Inter-annotator agreement (Cohen's κ) averaged 0.8193 (Almost Perfect; Table 5), with 311 of 1,000 instances (31.1%) disputed – concentrated in Humanity (120) and Sentiment (180). Disputes were resolved by the researcher via manual reconciliation against the operational definitions, without consulting the LLM prediction. Since aspect labels are binary, a conflict implies $\{E1,E2\} = \{0, 1\}$, so the LLM is guaranteed to match one expert, making 50% (not near 0%) the correct chance baseline; of 463 label-level conflicts (Table 4), the final decision matched the LLM in only 26.6% of cases (aspect-only: 31.4%), well below this baseline, confirming the LLM had no influence on reconciliation. The reconciled label matched Expert 2 in 59.6% of conflicts, plausibly reflecting more explicit justification notes from Expert 2, though this is interpretive rather than independently verified.

Table 4: Manual Reconciliation Outcomes for Disputed Labels ($n = 311$ instances, 463 label-level conflicts)

Label	N Conflict	Final=LLM	Final=E1	Final=E2	Final=Neither
Deity	17	4 (23.5%)	12 (70.6%)	5 (29.4%)	0
Humanity	120	39 (32.5%)	61 (50.8%)	59 (49.2%)	0
Unity	29	10 (34.5%)	16 (55.2%)	13 (44.8%)	0
Democracy	75	17 (22.7%)	21 (28.0%)	54 (72.0%)	0
Social Justice	42	19 (45.2%)	19 (45.2%)	23 (54.8%)	0
Sentiment	180	34 (18.9%)	50 (27.8%)	122 (67.8%)	8 (4.4%)
Total	**463**	**123 (26.6%)**	**179 (38.7%)**	**276 (59.6%)**	**8 (1.7%)**

Agreement between the reconciled ground truth and the original LLM labels was $\kappa = 0.8092$ overall (Almost Perfect for 4 of 6 dimensions; Table 5), supporting the pseudo-labeled dataset's reliability for training. Humanity showed the weakest LLM-expert agreement ($\kappa = 0.6519$), driven by 150 false positives where the LLM over-labeled general negativity as Humanity violations a known limitation of zero-shot LLM annotation [17] and its operational definition was refined accordingly. Because training and in-pipeline test data share the same annotation process, this expert set also serves as an independent, secondary evaluation set (Section 3); it is not fully out-of-distribution, being drawn from the same four sources and modest in size ($n = 1{,}000$) relative to the 41,138-instance training pool, and all 12 configurations share a single deterministic split (Table 7) without repeated-split (k-fold) validation both limitations are revisited in Section 4.

Table 5: Inter-Annotator ($n = 1{,}000$) and LLM-Expert (κ) Agreement

Label	Expert-Expert κ	LLM-Expert κ
Deity	0.8913	0.8732
Humanity	0.7510	0.6519
Unity	0.8489	0.8762
Democracy	0.8285	0.8331
Social Justice	0.8739	0.8972
Sentiment	0.7220	0.7239
Average	**0.8193**	**0.8092**

2.4. Class Imbalance, Training, and Splitting

Aspect labels ranged from 8.27% (Divinity) to 66.09% (Humanity) positive; loss weighting (pos_weight in BCEWithLogitsLoss for aspects; class weights in CrossEntropyLoss for sentiment; Table 6) mitigated imbalance, held fixed across all 12 configurations. Models used: IndoBERTweet [18], IndoBERT [19], mBERT [20], and XLM-RoBERTa [21]. LoRA/QLoRA used rank $r = 16$, $\alpha = 32$, dropout 0.1 on query/key/value/dense modules (2,678,784 trainable parameters, $\leq$ 2.4% of any base model; Table 7); QLoRA additionally applied 4-bit NF4 quantization to base weights. The 41,138-instance dataset was split 70:15:15 (28,796/6,171/6,171) via stratified (on sentiment) `train_test_split` (seed = 42). Training used AdamW (weight_decay = 0.01), constant learning rate 5×10^{-5}, batch size 32, max 10 epochs, and early stopping (patience = 3) requiring both validation-F1 improvement and non-degrading validation loss; 10 of 12 experiments triggered early stopping (mBERT_qlora and XLM-RoBERTa_qlora trained the full 10 epochs). The split seed was fixed, but the training loop itself (weight initialization, batch shuffling) was not seeded – disclosed as a reproducibility limitation revisited in Section 4.

Table 6: Loss-Weighting Values for Class-Imbalance Handling

Aspect (pos_weight)	**Value**
Deity	11.0890
Humanity	0.5131
Unity	7.7340
Democracy	1.3159
Social Justice	4.5655
Sentiment (class weight)	**Value**
Negative	0.7679
Neutral	0.9862
Positive	1.4625

Table 7: Trainable Parameters (LoRA/QLoRA) by Base Model

Base Model	**Trainable**	**Total**	**%**
IndoBERT	2,678,784	127,120,128	2.11
IndoBERTweet	2,678,784	113,236,992	2.37
mBERT	2,678,784	180,532,224	1.48
XLM-RoBERTa	2,678,784	280,722,432	0.95

2.5. Evaluation

Performance was measured via macro-averaged Precision, Recall, and F1-Score [22] (Eq. 1-3), chosen for robustness to class imbalance.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = \frac{2\,Precision \cdot Recall}{Precision + Recall} \tag{3}$$

Because all 12 configurations were evaluated on one deterministic split, 95% confidence intervals for aspect, sentiment, and average F1 were additionally estimated via non-parametric bootstrap resampling (2,000 resamples) on both the in-pipeline test set ($n = 6,171$) and the expert ground truth ($n = 1,000$); formal significance tests (paired t-test, McNemar's) were not applied, as they presuppose repeated/independent sampling [23]. All experiments used HuggingFace Transformers/PEFT on an NVIDIA RTX 5060 (8 GB VRAM), 32 GB RAM, AMD Ryzen AI 9 HX 370. Table 8 summarizes the full hyperparameter configuration for replication.

Table 8: Full Hyperparameter and Training Configuration

Parameter	**Value**
Tokenizer	WordPiece (IndoBERT, IndoBERTweet, mBERT); SentencePiece (XLM-RoBERTa)
Max Sequence Length	128 tokens
Optimizer	AdamW (weight_decay = 0.01)
Learning Rate	5×10^{-5}, constant (no scheduler/warmup)
Batch Size	32
Max Epochs	10 (early-stopping patience = 3)
Early Stopping Criterion	Best validation F1 AND validation loss not worsening by >1%
LoRA Rank / Alpha / Dropout	$r = 16$ / $\alpha = 32$ / 0.1, on query/key/value/dense modules
QLoRA Quantization	4-bit NF4
Train / Val / Test Split	70% / 15% / 15% (stratified by sentiment, seed = 42)
Training-Loop Random Seed	Not fixed (weight init. and batch shuffling unseeded)

3. RESULT AND ANALYSIS

All 12 configurations were trained on 28,796 instances and evaluated on 6,171 held-out test instances plus the independent 1,000-instance expert ground truth. Because the corpus is TikTok-dominated (63% of raw data, with a distinct short-form, reactive discourse register), generalization to other platforms or regional-language-heavy text should not be assumed without further platform-stratified evaluation. Data/code sharing is restricted by platform terms of service, but all hyperparameters, loss weights, split procedure, and formulas needed for replication are reported in Section 2; the Kaggle and NusaX subsets are independently public.

3.1. Training Behavior

FFT models showed decreasing training loss but rising validation loss in later epochs (overfitting; 10/12 configurations triggered early stopping), whereas LoRA/QLoRA produced more stable validation curves, consistent with adapter-based fine-tuning mitigating overfitting [11]. IndoBERT and IndoBERTweet generally converged better than the multilingual models.

3.2. Aspect and Sentiment Classification

Table 9 reports in-pipeline test results. Aspect F1 ranged 0.70–0.76; the best was XLM-RoBERTa LoRA (0.7579), closely followed by IndoBERT LoRA (0.7562) and IndoBERTweet FFT (0.7527) - multilingual pretraining is not inherently disadvantageous once adapter-based fine-tuning is applied. mBERT underperformed consistently across all three strategies (0.7139-0.7172), broadly consistent with a multilingual-capacity trade-off reported elsewhere [21], [8], though XLM-RoBERTa's stronger showing suggests this is model-specific; alignment between pretraining corpus and target-language characteristics [7, 24] likely matters more for some architectures than others. For sentiment, XLM-RoBERTa QLoRA (0.7770), IndoBERTweet FFT (0.7774), and IndoBERT LoRA (0.7750) led; mBERT again trailed, consistent with sentiment polarity carrying more explicit linguistic cues than fine-grained aspect violations [9], [23]. Averaged per strategy (Table 10), both LoRA (0.7475) and QLoRA (0.7480) outperformed FFT (0.7311) on Average F1 despite using only 1–2% of trainable parameters [13], [11] - a favorable trade-off given the storage/inference savings shown below.

Table 9: Aspect and Sentiment F1 on the In-Pipeline Test Set ($n = 6{,}171$)

Model	Strategy	Aspect F1	Sentiment F1	Avg. F1
IndoBERT	FFT	0.6982	0.7490	0.7236
	LoRA	0.7562	0.7750	**0.7656**
	QLoRA	0.7456	0.7677	0.7567
IndoBERTweet	FFT	0.7527	0.7774	0.7650
	LoRA	0.7378	0.7695	0.7537
	QLoRA	0.7427	0.7731	0.7579
mBERT	FFT	0.7161	0.6978	0.7070
	LoRA	0.7139	0.6981	0.7060
	QLoRA	0.7172	0.7129	0.7150
XLM-RoBERTa	FFT	0.7021	0.7558	0.7289
	LoRA	**0.7579**	0.7715	0.7647
	QLoRA	0.7474	**0.7770**	0.7622

Table 10: Average Performance per Fine-Tuning Strategy (Across 4 Base Models)

Strategy	Avg. Aspect F1	Avg. Sentiment F1	Average F1
FFT	0.7173	0.7450	0.7311
LoRA	**0.7414**	**0.7535**	**0.7475**
QLoRA	0.7382	**0.7577**	0.7480

Table 11 reports per-category aspect F1 against ground truth for all 12 configurations. Two patterns are consistent across every configuration without exception: Humanity achieves the highest or near-highest F1 (0.7833-0.8212), reflecting its dominant support (565 of 1,272 aggregate positive GT labels), while Unity (Unity) is the weakest category in *every* configuration (F1 range 0.4976-0.6667; Table 12), driven by systematic over-prediction: Recall exceeds Precision by 0.06–0.57, most severely for IndoBERT FFT (Precision = 0.34 at Recall = 0.91) - consistent with cultural/contextual understanding playing a critical role in value-laden Indonesian text [25] and with semantic overlap between Unity and the much more frequent Humanity category. XLM-RoBERTa LoRA shows both the highest Unity Precision (0.6371) and smallest Recall-Precision gap (0.0620), and correspondingly

the highest Unity F1 (0.6667). FFT consistently shows the most severe Unity over-prediction across all four base models (gap 0.30-0.57), suggesting full-parameter updates more readily learn spurious Unity-Humanity correlations than adapter-constrained fine-tuning. For sentiment, Neutral is consistently weakest on both the test set (F1 0.6454-0.7567) and ground truth (F1 0.4822-0.5583), driven by low Recall, consistent with its smaller training support relative to Negative (9,733 vs. 12,500 instances).

Table 11: Per-Category Aspect F1 Against Ground Truth, All 12 Configurations (n_{GT} = 1,000; Support: Deity = 88, Humanity = 565, Unity = 113, Democracy = 299, Social Justice = 207)

Model	Strategy	Deity	Humanity	Unity	Democracy	Social Justice
IndoBERT	FFT	0.7064	0.8117	0.4976	0.7670	0.6654
	LoRA	0.7535	0.8110	0.6505	0.7273	0.7413
	QLoRA	0.7257	0.8085	0.6389	0.7520	0.6787
IndoBERTweet	FFT	**0.8351**	0.8212	0.6467	0.7426	**0.7197**
	LoRA	0.7568	**0.8212**	0.6312	0.7561	0.6765
	QLoRA	0.7921	0.8154	0.6426	0.7657	0.6840
mBERT	FFT	0.8022	0.7969	0.6323	**0.7695**	0.5663
	LoRA	0.7805	0.8047	0.6241	0.7532	0.6412
	QLoRA	0.7624	0.8035	0.6250	0.7348	0.6512
XLM-RoBERTa	FFT	0.7225	0.7833	0.5534	0.7455	0.6784
	LoRA	0.7751	0.8100	**0.6667**	0.7633	0.6572
	QLoRA	0.7222	0.7948	0.6548	0.7654	0.6903

Table 12: Precision/Recall for (Unity) vs. Ground Truth, All 12 Configurations (n_{GT} = 1,000, Support = 113)

Model	Strategy	Precision	Recall	R−P
IndoBERT	FFT	0.3422	0.9115	0.5693
	LoRA	0.5341	0.8319	0.2978
	QLoRA	0.5257	0.8142	0.2885
IndoBERTweet	FFT	0.5187	0.8584	0.3397
	LoRA	0.5053	0.8407	0.3354
	QLoRA	0.5427	0.7876	0.2449
mBERT	FFT	0.5169	0.8142	0.2973
	LoRA	0.5207	0.7788	0.2581
	QLoRA	0.5346	0.7522	0.2176
XLM-RoBERTa	FFT	0.4008	0.8938	0.4930
	LoRA	**0.6371**	0.6991	**0.0620**
	QLoRA	0.5476	0.8142	0.2666

3.3. Independent Evaluation Against Ground Truth

Because training and in-pipeline test labels share the same zero-shot LLM origin, in-pipeline scores may overstate real generalization. All 12 checkpoints were therefore also evaluated on the independent 1,000-instance expert ground truth (Table 13). Every configuration declined relative to the test set (average relative drop 6.71%, range 3.16–8.99%), confirming a genuine evaluation-circularity effect; in-pipeline F1 should be read as an upper bound. Critically, ranking changes: IndoBERT LoRA, the best in-pipeline configuration (0.7656), drops to a GT Average F1 of 0.6968 (largest relative drop, 8.99%), while IndoBERTweet QLoRA achieves the highest point-estimate GT Average F1 (0.7137) despite a modest in-pipeline rank, using only 10.27 MB of adapter storage. Bootstrap 95% CIs (2,000 resamples) show the top 3-4 configurations overlap substantially on both sets, so IndoBERTweet QLoRA's lead is best read as the top of an indistinguishable cluster rather than a confirmed single best model; its advantage over the weakest configurations (mBERT_lora, XLM-RoBERTa_base, IndoBERT_base), however, is statistically supported. GT-based CIs are roughly 2.45× wider than test-set CIs ($\approx \sqrt{6{,}171/1{,}000}$), as expected from sample size.

Table 13: In-Pipeline vs. Independent Ground-Truth Performance (n_{GT} = 1,000)

Model	Strategy	Test F1	GT F1	Gap	Drop (%)
IndoBERT	FFT	0.7236	0.6727	0.0509	7.04
	LoRA	0.7656	0.6968	0.0688	**8.99**
	QLoRA	0.7567	0.6954	0.0612	8.09

Model	Strategy	Test F1	GT F1	Gap	Drop (%)
IndoBERTweet	FFT	0.7650	0.7109	0.0541	7.07
	LoRA	0.7537	0.7062	0.0475	6.30
	QLoRA	0.7579	**0.7137**	0.0442	5.83
mBERT	FFT	0.7070	0.6846	0.0224	**3.16**
	LoRA	0.7060	0.6706	0.0354	5.02
	QLoRA	0.7150	0.6788	0.0362	5.07
XLM-RoBERTa	FFT	0.7289	0.6720	0.0569	7.80
	LoRA	0.7647	0.7057	0.0589	7.71
	QLoRA	0.7622	0.6980	0.0643	8.43

3.4. Computational Resource Evaluation

Table 14 summarizes checkpoint size, training time, and inference cost. All LoRA/QLoRA checkpoints are 10.27 MB versus 421.85-1,060.76 MB for FFT (97-99% storage reduction) and need substantially less inference VRAM (487-1,419 MB vs. 1,976-4,570 MB), though LoRA/QLoRA training time is longer (36.5-89.9 min vs. 27.4-58.5 min) since adapter methods needed more epochs before early stopping despite more stable validation curves. Combined with Table 10, this is a strictly favorable trade-off: LoRA/QLoRA match or exceed FFT's average F1 while using roughly 1-2% of trainable parameters and a fraction of storage/inference memory - directly relevant given IndoBERTweet QLoRA, the best ground-truth generalizer, needs only 10.27 MB.

Table 14: Computational Resource Usage per Experiment

Model	Strategy	Ckpt (MB)	Train (min)	Infer. VRAM (MB)	Infer. (sps)
IndoBERT	FFT	474.81	27.4	2,140	199.1
	LoRA	10.27	49.7	792	164.0
	QLoRA	10.27	36.5	487	256.1
IndoBERTweet	FFT	421.85	33.3	1,976	198.0
	LoRA	10.27	50.1	738	164.2
	QLoRA	10.27	38.2	434	252.2
mBERT	FFT	678.56	35.4	3,001	193.8
	LoRA	10.27	60.6	1,037	105.9
	QLoRA	10.27	73.6	733	139.3
XLM-RoBERTa	FFT	1,060.76	58.5	4,570	145.9
	LoRA	10.27	89.9	1,419	151.9
	QLoRA	10.27	58.5	1,115	140.1

3.5. Deployment Implications

Given the residual evaluation-circularity gap and the Unity/Neutral weaknesses already visible at the default 0.5 threshold, fully automated, human-out-of-the-loop deployment is not recommended. A human-in-the-loop workflow is proposed, where predictions flag candidate instances for moderator review rather than action removal directly, with per-label thresholds tuned (e.g. via precision-recall curves) to prioritize higher recall for Unity and Neutral given their higher real-world cost of missed detection; the raw probability outputs needed to construct these curves were not yet available at the time of writing and are left for follow-up work, alongside a targeted-sampling expansion of the ground-truth set and label-dependency modeling to address Unity's persistent weakness.

4. CONCLUSION

This study proposed a Pancasila-based ACSA framework for detecting negative content in Indonesian code-mixed social media, using a pseudo-labeled dataset (41,138 instances, confidence ≥ 0.75) built via zero-shot Llama 3 8B annotation and validated by two experts (inter-annotator $\kappa = 0.8193$; LLM-expert $\kappa = 0.8092$, both Almost Perfect). Across 12 configurations, Indonesian-specific pretraining did not confer a uniform advantage: IndoBERT and XLM-RoBERTa under LoRA achieved the top in-pipeline aspect F1-Scores, while mBERT underperformed consistently. LoRA and QLoRA outperformed Full Fine-Tuning on average (0.7475/0.7480 vs. 0.7311 Average F1) while cutting checkpoint size by up to 99% (10.27 MB vs. 421.85–1,060.76 MB), confirming the practical viability of parameter-efficient fine-tuning for resource-constrained Indonesian NLP.

Independent evaluation against expert-validated ground truth (n = 1,000, no overlap with training data) revealed a consistent 3.16–8.99% relative performance drop across all configurations (average 6.71%), confirming genuine evaluation circularity and revising the headline ranking: IndoBERTweet QLoRA achieved the best point-estimate ground-truth generalization (F1 = 0.7137, 10.27 MB), though bootstrap 95% CIs show this is not statistically distinguishable from the next three configurations (IndoBERTweet FFT/LoRA, XLM-RoBERTa LoRA) - best read as a top cluster rather than a single winner - while its advantage over the weakest configurations is statistically supported.

Contributions include: (1) a culturally grounded ACSA framework anchored in Pancasila principles; (2) a large-scale pseudo-labeled Indonesian code-mixed dataset validated through two-annotator agreement analysis; (3) a reproducible comparison of Transformer architectures and fine-tuning strategies; and (4) an independent ground-truth evaluation quantifying the gap between pseudo-labeled and human-judged performance. Key limitations: reliance on a single deterministic data split (mitigated, not replaced, by bootstrap CIs); a modest (n = 1,000), non-out-of-distribution ground-truth sample; an unseeded training loop; and a TikTok-dominated, code-mixing-narrow corpus that limits cross-platform/cross-dialect generalization claims. Future work should expand ground-truth coverage (targeted low-confidence and platform-stratified sampling), add repeated-split (k-fold) validation, quantify training-seed variance, and pursue data augmentation or label-dependency modeling to address the persistent weakness of the Unity category. Overall, combining culturally grounded aspect categories with parameter-efficient Transformer fine-tuning offers a promising, resource-feasible foundation for interpretable, value-sensitive content moderation in Indonesia, provided performance claims are grounded in independent human validation rather than pseudo-labeled metrics alone.

REFERENCES

[1] W. A. S. Indonesia, "Digital 2025 – your ultimate guide to the evolving digital world," 1 2025. [Online]. Available: https://wearesocial.com/id/blog/2025/01/digital-2025

[2] Q. A. Xu, V. Chang, and C. Jayne, "A systematic review of social media-based sentiment analysis: Emerging trends and challenges," *Decision Analytics Journal*, vol. 3, p. 100073, 6 2022.

[3] M. Ali, M. Hassan, K. Kifayat, J. Y. Kim, S. Hakak, and M. K. Khan, "Social media content classification and community detection using deep learning and graph analytics," *Technological Forecasting and Social Change*, vol. 188, 3 2023.

[4] P. R. Indonesia, "Undang-undang republik indonesia nomor 19 tahun 2016 tentang perubahan atas undang-undang nomor 11 tahun 2008 tentang informasi dan transaksi elektronik," Kementerian Sekretariat Negara Republik Indonesia, Tech. Rep., 11 2016. [Online]. Available: https://peraturan.bpk.go.id/Details/37582/uu-no-19-tahun-2016

[5] C. Tho, Y. Heryadi, I. H. Kartowisastro, and W. Budiharto, "A comparison of lexicon-based and transformer-based sentiment analysis on code-mixed of low-resource languages," in *Proceedings of 2021 1st International Conference on Computer Science and Artificial Intelligence, ICCSAI 2021*. Institute of Electrical and Electronics Engineers Inc., 2021, pp. 81–85.

[6] M. K. Nazir, C. N. Faisal, M. A. Habib, and H. Ahmad, "Leveraging multilingual transformer for multiclass sentiment analysis in code-mixed data of low-resource languages," *IEEE Access*, vol. 13, pp. 7538–7554, 2025.

[7] A. F. Hidayatullah, "Code-mixed sentiment analysis on indonesian-javanese-english text using transformer models," in *2024 8th International Conference on Information Technology, Information Systems and Electrical Engineering, ICITISEE 2024*. Institute of Electrical and Electronics Engineers Inc., 2024, pp. 340–345.

[8] A. N. Azhar and M. L. Khodra, "Fine-tuning pretrained multilingual bert model for indonesian aspect-based sentiment analysis," in *2020 7th International Conference on Advanced Informatics: Concepts, Theory and Applications, ICAICTA 2020*. Institute of Electrical and Electronics Engineers Inc., 9 2020.

[9] W. Liao, B. Zeng, X. Yin, and P. Wei, "An improved aspect-category sentiment analysis model for text sentiment analysis based on roberta," *Applied Intelligence*, vol. 51, pp. 3522–3533, 6 2021.

[10] M. A. Almasre, "Enhance the aspect category detection in arabic language using arabert and text augmentation," in *Proceedings of 2022 5th National Conference of Saudi Computers Colleges, NCCC 2022*. Institute of Electrical and Electronics Engineers Inc., 2022, pp. 7–10.

[11] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," in *The Tenth International Conference on Learning Representations*. OpenReview.net, 5 2022. [Online]. Available: http://arxiv.org/abs/2106.09685

[12] S. S. Alahmari, L. O. Hall, P. R. Mouton, and D. B. Goldgof, "Repeatability of fine-tuning large language models illustrated using qlora," *IEEE Access*, vol. 12, pp. 153 221–153 231, 2024.

[13] S. Joshi, M. S. Khan, A. Dafe, K. Singh, V. Zope, and T. Jhamtani, "Fine tuning llms for low resource languages," in *Proceedings - 2024 5th International Conference on Image Processing and Capsule Networks, ICIPCN 2024*. Institute of Electrical and Electronics Engineers Inc., 2024, pp. 511–519.

[14] A. M. M. Nasoha, A. N. Atqiya, I. K. Anam, H. A. Natasya, and R. Salsabila, "Implementasi nilai - nilai pancasila dalam penggunaan media sosial," *JURNAL HUKUM, POLITIK DAN ILMU SOSIAL*, vol. 4, pp. 163–174, 3 2025.

[15] G. I. Winata, A. F. Aji, S. Cahyawijaya, R. Mahendra, F. Koto, A. Romadhony, K. Kurniawan, D. Moeljadi, R. E. Prasojo, P. Fung, T. Baldwin, J. H. Lau, R. Sennrich, and S. Ruder, "NusaX: Multilingual parallel sentiment dataset for 10 Indonesian local languages," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2023, pp. 815–834.

[16] V. Nguyen and C. Pham, "Leveraging large language models for suicide detection on social media with limited labels," in *Proceedings - 2024 IEEE International Conference on Big Data, BigData 2024*. Institute of Electrical and Electronics Engineers Inc., 2024, pp. 8550–8559.

[17] Z. Ye, Y. Geng, J. Chen, X. Xu, S. Zheng, F. Wang, J. Chen, J. Zhang, and H. Chen, "Zero-shot text classification via reinforced self-training," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 7 2020, pp. 3014–3024. [Online]. Available: https://aclanthology.org/2020.acl-main.272

[18] A. N. Hakim, Y. Sibaroni, and S. S. Prasetyowati, "Detection of hate-speech text on indonesian twitter social media using indobertweet-bilstm-cnn," in *2024 12th International Conference on Information and Communication Technology, ICoICT 2024*. Institute of Electrical and Electronics Engineers Inc., 2024, pp. 374–381.

[19] N. Alamsyah, E. Noersasongko, G. F. Shidik, and N. Rijati, "Fine-grained sentiment classification of public opinion on electric cars in indonesia using indobert," in *Proceedings - 2024 International of Seminar on Application for Technology of Information and Communication: Smart And Emerging Technology for a Better Life, iSemantic 2024*. Institute of Electrical and Electronics Engineers Inc., 2024, pp. 502–508.

[20] B. K. Jha, C. M. V. S. Akana, and R. Anand, "Question answering system with indic multilingual-bert," in *Proceedings - 5th International Conference on Computing Methodologies and Communication, ICCMC 2021*. Institute of Electrical and Electronics Engineers Inc., 4 2021, pp. 1631–1638.

[21] P. Ferdosian, S. Grace, V. Manikandan, L. Moles, D. Datta, and D. Brown, "Improving the efficiency and effectiveness of multilingual classification methods for sentiment analysis," in *2021 IEEE Systems and Information Engineering Design Symposium, SIEDS 2021*. Institute of Electrical and Electronics Engineers Inc., 4 2021.

[22] M. P. Geetha and D. K. Renuka, "Improving the performance of aspect based sentiment analysis using fine-tuned bert base uncased model," *International Journal of Intelligent Networks*, vol. 2, pp. 64–69, 1 2021.

[23] J. Šmíd and P. Král, "Cross-lingual aspect-based sentiment analysis: A survey on tasks, approaches, and challenges," *Information Fusion*, vol. 120, 8 2025.

[24] A. F. Hidayatullah, K. Kalinaki, M. M. Aslam, R. Y. Zakari, and W. Shafik, "Fine-tuning bert-based models for negative content identification on indonesian tweets," in *ICITDA 2023 - Proceedings of the 2023 8th International Conference on Information Technology and Digital Applications*. Institute of Electrical and Electronics Engineers Inc., 2023.

[25] R. V. Lucky, B. A. Herkanyaka, V. F. P. Basuki, I. H. Parmonangan, and Diana, "Unveiling sentiments in javanese text: A study on sentiment analysis for the javanese language," in *Proceeding - IEEE 9th Information Technology International Seminar, ITIS 2023*. Institute of Electrical and Electronics Engineers Inc., 2023.