Sistemasi: Jurnal Sistem Informasi
Volume 15, Nomor 10, 2026: 4260-4273
ISSN:2302-8149
e-ISSN:2540-9719

Benchmarking IndoBERT and IndoBERTweet on Emotion and Sentiment of Indonesian App Reviews

[1]Natalie Grace Widjaja Kuswanto, [2]Rinabi Tanamal*
[1,2]Program Studi Sistem Informasi, Fakultas Teknologi Informasi, Universitas Ciputra Surabaya
[1,2]CitraLand CBD Boulevard, Surabaya, Jawa Timur, 60219, Indonesia
*e-mail: r.tanamal@ciputra.ac.id

(***received*****:** 5 October 2026, ***revised*****:** 14 October 2026, ***accepted*****:** 16 October 2026)

Abstract

Indonesian mobile application reviews are informal free-text feedback. Their sentiment and emotion labels are costly to obtain and only moderately reliable. This study determines which pre-trained encoder, IndoBERT or IndoBERTweet, classifies emotion and sentiment more reliably in Indonesian app reviews, and tests whether combining or routing between the two encoders improves on the reference single model. IndoBERT was pre-trained on general Indonesian text and IndoBERTweet on Indonesian tweets. Both were fine-tuned for six-class emotion and three-class sentiment classification on 21,065 de-duplicated reviews. Five random seeds, McNemar tests, and paired bootstrap confidence intervals evaluated confidence-based and task-based routing, probability averaging, sentiment derived from emotion, and a lexical register analysis. IndoBERTweet outperformed IndoBERT on emotion, reaching a macro-F1 of 0.627 against 0.596, with the advantage concentrated on reviews containing slang. Because IndoBERTweet performed at least as well on both formal and informal reviews, confidence-based routing gave no additional gain. Averaging the probabilities of ten models raised emotion macro-F1 significantly to 0.652, an improvement of 0.024 over the reference single model. Sentiment derived from the predicted emotion was never worse than a dedicated sentiment head. It performed significantly better for a single IndoBERTweet model, reaching 0.807 against 0.796, eliminated 10.2 to 14.6% of reviews with contradictory separate heads, and halved inference cost. Single-model modifications that did not degrade performance converged near 0.62 to 0.63, consistent with annotation reliability limiting attainable performance. These findings give practitioners an evidence-based default configuration: IndoBERTweet as the encoder, with sentiment read directly from the predicted emotion label.

Keywords: *annotation agreement, emotion classification, IndoBERTweet, model routing, sentiment analysis.*

1 Introduction

Application stores have turned user feedback into a continuously growing body of unsolicited free-text data that developers can mine for requirements, complaints, and satisfaction signals [1]. In Indonesia, users' perception of applications has typically been studied with questionnaires. Technology-acceptance analyses of Indonesian application adoption, for example, related the intention to use or adopt an application to predefined constructs such as perceived usefulness, ease of use, trust, and habit, as shown for Gojek users through the UTAUT2 model [2] and for TikTok Shop users through the Technology Acceptance Model [3]. App reviews are a different kind of qualitative evidence: they are written spontaneously in the user's own words, and they often disagree with the star rating attached to them, as shown for the PLN Mobile application [4]. While the unstructured nature of these reviews provides valuable insights into the experience and perceptions of users, their textual content must first be systematically classified to identify and quantify the underlying sentiments and opinions.

Sentiment and emotion annotation provide a way to transform unstructured user reviews into analyzable information. Sentiment analysis assigns polarity, whereas emotion analysis provides a finer-grained classification, such as anger, sadness, or love, allowing different emotional responses within the same negative polarity to be distinguished. Studies on Indonesian app reviews have mainly focused on sentiment classification using IndoBERT [4], [5], [6] while emotion classification has shown lower and more variable performance, particularly with confusion between happiness and

 ISSN:2302-8149
e-ISSN:2540-9719

love [7], [8]. The dataset by Riccosan and Saputra is one of the few Indonesian resources that annotates both sentiment and emotion, with three sentiment classes and six emotion classes [9]. Its annotators achieved Cohen's kappa values of 0.604 for emotion and 0.625 for sentiment, indicating moderate to substantial agreement [9], [10].

Reviews in this dataset contain a mix of standard Indonesian with slang, abbreviations, and code-switching. IndoBERT was pre-trained on general Indonesian text [11], meanwhile IndoBERTweet was trained on Indonesian tweets with a vocabulary adapted to social media language [12]. IndoBERTweet has also been reported to match or outperform IndoBERT in sentence-level sentiment classification of Gojek reviews, depending on the validation scheme used [13]. This suggests a natural design: if IndoBERT is better on formal reviews and IndoBERTweet on informal ones, a router that selects a model for each review or each task should beat either model alone. However, this approach has not yet been evaluated for joint sentiment and emotion classification on the same Indonesian app review dataset, nor has the performance of combining both models through a simpler averaging strategy been examined. In addition, the extent to which classification errors arise from model limitations or annotation inconsistencies remains unclear.

This study therefore asks four questions. (i) How do IndoBERT and IndoBERTweet compare on emotion and sentiment classification of Indonesian app reviews, and does the difference depend on the register of the review? (ii) Does routing between the two models, by confidence or by task, improve on the better single model? (iii) Does averaging the probabilities of several models help, and at what inference cost? (iv) Can sentiment be derived from the predicted emotion instead of being predicted by a dedicated head, given that the dataset defines sentiment as a grouping of emotion classes? Answering these questions on a de-duplicated, seed-controlled protocol with confidence intervals gives developers a defensible default encoder and shows how much of the remaining error may be attributed to label reliability. The following sections present the related work, dataset and methodology, experimental results and discussion, and conclusions.

2 Literature Review

Research on Indonesian sentiment classification of application reviews has centred on IndoBERT. Asri et al. labelled 1,000 PLN Mobile reviews with the InSet lexicon and reached 81% accuracy with IndoBERT [4]. Tarwoto et al. reported an average accuracy of 97.28% on three-class sentiment of Mobile JKN reviews [5], and Pradhisa and Fajriyah obtained 89.84% for Shopee and 88.12% for Bukalapak reviews, with reference labels generated by the Vader lexicon [6]. Beyond app stores, Jayadianti et al. found that adding a recurrent convolutional network to IndoBERT improved on IndoBERT alone, reaching an F1-score of 93.27% on the IndoNLU sentiment data [14], and a study of Gemini reviews published in this journal obtained 94.57% accuracy with Naïve Bayes on two polarity classes [15]. These figures are high, but they share features that limit their transfer: two or three classes, labels produced by lexicons in several cases, samples of a few thousand reviews from one or two applications, and IndoBERT as the only encoder. None examines how a model behaves when formal and informal reviews coexist, and none reports how reliable the labels are.

Whether a model pre-trained on informal text should be preferred has been examined mainly outside emotion classification. IndoBERTweet was pre-trained on Indonesian tweets, with a vocabulary initialized for social-media language, and outperformed IndoBERT, mBERT, and MalayBERT on several Indonesian Twitter benchmarks [12]. On Gojek reviews segmented into sentences, IndoBERTweet reached a macro-F1 of 0.729 against 0.714 for IndoBERT under stratified five-fold validation, although the two models tied at 0.701 under a time-based split [13]. Comparisons between IndoBERT and multilingual BERT point in the opposite direction on other tasks: on a balanced binary task of gambling-comment detection, IndoBERT reached 98% accuracy against 96% for mBERT, a difference confirmed by McNemar's test [16], although the two-class formulation and the balanced label distribution limit how directly this figure can be compared to the six-class emotion task addressed here. IndoBERT was also the best of the three encoders tested on this dataset in a multilabel formulation that predicts sentiment and emotion jointly (micro-F1 0.77 versus 0.72 for mBERT) [17]. Hybrid designs that combine two encoders have also been reported, such as IndoBERT features with an mBERT classifier for multilabel toxic-comment detection (F1 0.9032) [18]. Taken together, these studies show that the choice of pre-training corpus matters, but they compare encoders

 ISSN:2302-8149
e-ISSN:2540-9719

one task and one dataset at a time and leave the choice of encoder to the researcher rather than to the system.

Emotion classification has received less attention and reports lower scores. Rizky and Hidayat fine-tuned IndoBERT on 5,079 tweets with five emotion classes and observed signs of overfitting [8]. Saputra et al. reached 73% accuracy on six emotions from social media and noted that love was frequently confused with happiness [7]. Christian et al. compared IndoBERT and DistilBERT on e-commerce reviews with augmentation and reported that bagging five IndoBERT models, at 79.77% accuracy, did not exceed the best single configuration of about 80% [19]. On the dataset used here, Cahya et al. applied SMOTE and text augmentation to the emotion labels and reported an F1-score of 0.86, against 0.62 on the original data [20]. Their pipeline created the balanced corpus before splitting it into training, validation, and test sets, and every class contained 7,307 samples, so synthetic near-duplicates of training samples can appear in the test partition; the original-data figure is therefore the appropriate reference. Emotion results are thus lower, less consistent, and rarely accompanied by any statement on how reliable the labels are, and the evidence on combining models is limited to one negative result [19] and one hybrid design [18].

Human agreement is the natural yardstick for labels of this kind. Cohen's kappa corrects observed agreement for chance [21], and the scale of Landis and Koch labels 0.41 to 0.60 as moderate and 0.61 to 0.80 as substantial [10]. Artstein and Poesio observe that thresholds appropriate for computational linguistics have never been settled, although 0.67 is cited there as the accepted reference threshold [22]. Agreement in affective annotation is typically modest: Bobicev and Sokolova summarise studies with kappa between 0.26 and 0.75, report 0.46 for their own annotation of forum posts, and show that classifier F-score depends on how well the annotators agreed (0.794 on posts with unanimous labels against 0.571 on all posts) [23]. The dataset used here reports kappa of 0.604 for emotion and 0.625 for sentiment [9], and the Gojek study reports a Fleiss' kappa of 0.636 on a subset of its sentence-level annotations [13]. Confident learning offers a model-based complement, estimating which labels are likely wrong from out-of-sample predicted probabilities [24]. Averaging the predictions of several models is a long-established way of reducing variance [25], but whether it helps more than choosing a model per input has not been examined for Indonesian reviews.

Three gaps follow from this review. First, IndoBERT and IndoBERTweet have not been compared on both emotion and sentiment of the same app reviews with a mechanism that selects between the two encoders by task or by confidence rather than fixing one encoder in advance; the closest comparison [13] evaluated the two encoders directly on a single task and a different corpus, without such a selection mechanism, and the register-dependent hypothesis behind model selection has not been tested against a seed-controlled baseline. Second, sentiment and emotion are usually predicted by separate heads or models, although this dataset defines sentiment as a grouping of emotion classes [9], a relationship that Cahya et al. used to model emotion alone [20]; whether deriving sentiment from emotion matches a dedicated head has not been measured. Although the dataset paper [9], the companion multilabel study [17], and the SMOTE-based study [20] all use the same corpus, none of them investigates task-specific routing between encoders: the dataset paper [9] reports only annotation and baseline statistics, the multilabel study [17] evaluates encoders jointly on a fixed architecture, and the SMOTE-based study [20] addresses class imbalance within a single encoder rather than a choice between encoders. Third, label reliability is rarely connected to the performance ceiling. This study addresses these gaps by comparing the two encoders under a controlled protocol, testing routing and probability averaging against the best single model, measuring derived against dedicated sentiment, and relating a plateau of single-model results to annotation reliability. To the extent covered by the literature reviewed here, this combination has not been examined before, and it constitutes the contribution of the present study.

3 Research Method

The research process, summarized in Figure 1, comprised several stages: data acquisition and cleaning, stratified data splitting, fine-tuning of two Indonesian pre-trained encoders, development of the combination and hierarchical approaches, and evaluation using statistical tests, register analysis, ablation studies, and label-reliability diagnostics. All experiments were conducted on a single GPU,

 ISSN:2302-8149
e-ISSN:2540-9719

with all reported results generated using the same fixed training and evaluation pipeline to maintain consistency across tables.

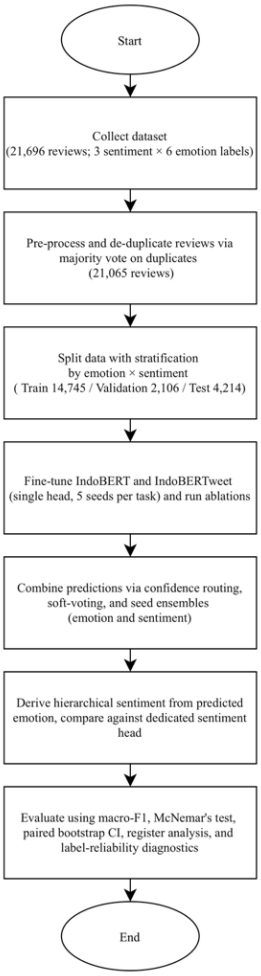


Figure 1 Research flow

Data collection and annotation. The dataset of Riccosan and Saputra was collected by scraping Google Play Store reviews of mobile applications in Indonesia, restricted to Indonesian-language reviews from Indonesian users [9]. Two annotators labelled each review with one of six emotions (anger, fear, sad, happy, love, neutral) using the word lists of an emotion hierarchy, and the sentiment label was assigned from the emotion: happy and love to positive; anger, sad, and fear to negative; and neutral to neutral. Announcements, notifications, random characters, and reviews without emotion words were labelled neutral [9]. This is the qualitative coding step of the dataset. The agreement between the two annotators was Cohen's kappa of 0.6044 for emotion and 0.6252 for sentiment [9]. The sentiment-from-emotion mapping was verified on the released file, and it holds for every row. After excluding the header row, 21,696 data rows remain, one fewer than the figure of 21,697 reported in the dataset paper [9].

Pre-processing and data split. URLs, user mentions, emojis, and punctuation were removed, text was lower-cased, elongated characters were shortened, numeric reduplication markers such as kata2 were expanded, and a dictionary of common slang forms was normalized. Sixteen rows that became empty after cleaning were removed, leaving 21,680 rows. Exact duplicate texts were then merged: among 346 groups of duplicates, 68 groups (19.7%) carried conflicting labels, which is a first indication of annotation noise, and the majority label of each group was kept. This left 21,065 reviews. Training, validation, and test partitions of 70%, 10%, and 20% were drawn with stratification on the joint emotion–sentiment label, and because duplicates were merged before splitting, no text

 ISSN:2302-8149
e-ISSN:2540-9719

appears in more than one partition. Table 1 gives the emotion distribution after de-duplication and the partition sizes; the imbalance ratio between the largest and smallest class is about 40:1.

Table 1 Emotion distribution after de-duplication and partition sizes

Emotion	Sentiment Group	Reviews	Share (%)
Neutral	Neutral	7,251	34.4
Happy	Positive	5,998	28.5
Sad	Negative	3,712	17.6
Anger	Negative	2,656	12.6
Fear	Negative	1,266	6.0
Love	Positive	182	0.9
Total		**21,065**	**100.0**
Train / Validation / Test		14,745 / 2,106 / 4,214	70 / 10 / 20

Models and training. Two BERT-base encoders [26] were fine-tuned; IndoBERT (indobenchmark/indobert-base-p1) [11] and IndoBERTweet (indolem/indobertweet-base-uncased) [12]. Each received a single classification head, consisting of dropout and a linear layer on the [CLS] representation, and was trained separately for the emotion task and for the sentiment task. Training used class-weighted cross-entropy, the AdamW optimizer, a linear schedule with 10% warm-up, and early stopping on validation macro-F1. Table 2 lists the hyperparameters, which were the same for both models. Every model-task pair was trained with five seeds (42 until 46), giving 20 models, so that means carry a standard deviation and ensembles can be built from different runs.

Table 2 Hyperparameters of the main protocol

Parameter	Value
Maximum sequence length	128 tokens
Batch size	16
Optimizer, learning rate, weight decay	AdamW, 2e-5, 0.01
Warm-up ratio	0.10 of total steps
Dropout	0.30
Maximum epochs / patience	15 / 4 (validation macro-F1)
Loss	Class-weighted cross-entropy
Pooling	[CLS]
Random seeds	42, 43, 44, 45, 46

Combination strategies. Let $p_{IB}(x)$ and $p_{IBT}(x)$ denote the class-probability vectors that IndoBERT and IndoBERTweet assign to a review x. Confidence fallback routing uses a primary model (IndoBERTweet for emotion, IndoBERT for sentiment, chosen on validation set) unless its maximum probability falls below a threshold t, in which case the secondary model is used; t was selected on the validation set. Task routing simply assigns each task to its better model. Soft-voting averages the two probability vectors as in Equation (1) and takes the arg-max, and a weighted variant uses a mixing weight tuned on the validation set. Seed ensembles average the probability vectors of the five seeds of one model or of all ten models of both encoders. The strategies that combine one IndoBERT and one IndoBERTweet model (routing, soft-voting, and weighted soft-voting) use a single pair of models.

$$p_{vote}(x) = \frac{1}{2}[p_{IB}(x) + p_{IBT}(x)] \quad (1)$$

Hierarchical sentiment. Because the dataset defines sentiment as a grouping of emotion classes, the sentiment of a review can be obtained from any predicted emotion through Equation (2). Derived sentiment was computed for every emotion model and combination and compared with a dedicated sentiment head trained on the same data. A dedicated head requires one forward pass per model for each task, so producing both labels costs twice as many passes as the derived scheme.

 ISSN:2302-8149
e-ISSN:2540-9719

$$\hat{y}_s = \begin{cases} Positive, if\ \hat{y}_e \in \{Happy, Love\} \\ Negative, if\ \hat{y}_e \in \{Anger, Sad, Fear\} \\ Neutral, otherwise \end{cases} \quad (2)$$

Register analysis. A review was labelled informal if it contains at least one entry of the slang dictionary used in pre-processing (1,739 test reviews) and formal otherwise (2,475 reviews). This is a lexical proxy for register, not a linguistic judgement.

Evaluation. The primary metric is macro-averaged F1, defined in Equation (3), where P_c and R_c are the precision and recall of class c, and K is the number of classes. It weights the six emotion classes equally and is therefore sensitive to minority classes fear and love; accuracy is reported alongside it. Differences between methods were tested with the McNemar test on paired correct and incorrect outcomes at a significance level of 0.05, and 95% confidence intervals were obtained by paired bootstrap resampling of the test set with 1,000 replicates [27]. All thresholds, mixing weights, and ensemble memberships were chosen on the validation set only.

$$F1_{macro} = \frac{1}{K}\sum_{c=1}^{K}\frac{2 \cdot P_c \cdot R_c}{P_c + R_c} \quad (3)$$

Label-reliability diagnostics and ablations. Agreement between annotators is expressed with Cohen's kappa, Equation (4) [21], where p_o is the observed agreement and p_e the agreement expected by chance. Duplicate texts with conflicting labels were counted during de-duplication, and confident learning with out-of-fold probabilities from five-fold cross-validation was used to flag training reviews whose labels are likely wrong [24].

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (4)$$

To test whether modelling choices could raise the emotion score, the following variants were trained with three seeds each on the same split, except where noted: minimal pre-processing that keeps emojis, punctuation, and slang; label smoothing; hyperparameters found by an Optuna search [28]; a large IndoBERT encoder (indobert-large-p1); XLM-RoBERTa large [29]; mean-max pooling with focal loss [30] and a higher learning rate for the head; and variants that remove, down-weight, or relabel training reviews flagged by confident learning. The mean–max pooling with focal loss variant was trained with five seeds per model/task (20 models in total) rather than three, so that its own ten-model ensemble (five IndoBERTweet plus five IndoBERT seeds) could be compared directly against the ten-model ensemble of the main protocol (M8) under the same seed count. Earlier screening on a preliminary split compared oversampling, undersampling, and text augmentation of the kind studied by Cahya et al. [20] on frozen [CLS] features with a small classifier, and five-class variants that merge love into happy. Multilingual BERT appeared only in the preliminary experiments and is not part of the final protocol.

4 Results and Analysis

Single models and combination strategies. Table 3 and Figure 2 report the two single models and the combination strategies on the original test labels. Across five seeds IndoBERTweet reached an emotion macro-F1 of 0.627 ± 0.003 compared with 0.596 ± 0.013 for IndoBERT. The seed-42 pair (M1 and M2) differs by 0.025, with a bootstrap interval of -0.054 to 0.004 that narrowly includes zero, so the case for the encoder difference rests mainly on the five-seed means, which are about three points apart, with seed standard deviations of 0.003 and 0.013. A two-sample t-test on the five-seed means confirmed that this difference is statistically significant (t = 5.20, df = 4.4, p = 0.005, Welch's correction for unequal variances), supporting the encoder-level comparison beyond the single seed-42 pair discussed above. For sentiment, the two seed-42 models were level (0.797 versus 0.796).

Table 3 Performance on the original test set (n = 4,214)

Method	Passes	Emotion macro-F1 [95% CI)	Emotion accuracy	Sentiment Macro-F1	Δ emotion vs M2 [95% CI]
M1 IndoBERT (seed 42)	1	0.603 [0.577, 0.626]	0.691	0.797	−0.025 [−0.054,

 ISSN:2302-8149
e-ISSN:2540-9719

					0.004]
M2 IndoBERTweet (seed 42)	1	0.628 [0.599, 0.655]	0.705	0.796	reference
M3 Confidence fallback routing	≈1.05	0.628 [0.598, 0.654]	0.708	0.804	−0.000 [−0.008, 0.006]
M4 Soft-voting IndoBERT + IndoBERTweet	2	0.638 [0.608, 0.663]	0.711	0.808	+0.009 [−0.005, 0.026]
M5 Weighted soft-voting	2	0.629 [0.600, 0.655]	0.709	0.808	+0.001 [−0.008, 0.012]
M6 Ensemble of 5 IndoBERT seeds	5	0.623 [0.596, 0.649]	0.710	0.811	−0.005 [−0.030, 0.021]
M7 Ensemble of 5 IndoBERTweet seeds	5	0.645 [0.617, 0.670]	0.720	0.811	+0.017 [0.000, 0.034]
M8 Ensemble of all 10 models	10	0.652 [0.624, 0.676]	0.729	0.814	+0.024 [0.005, 0.044]

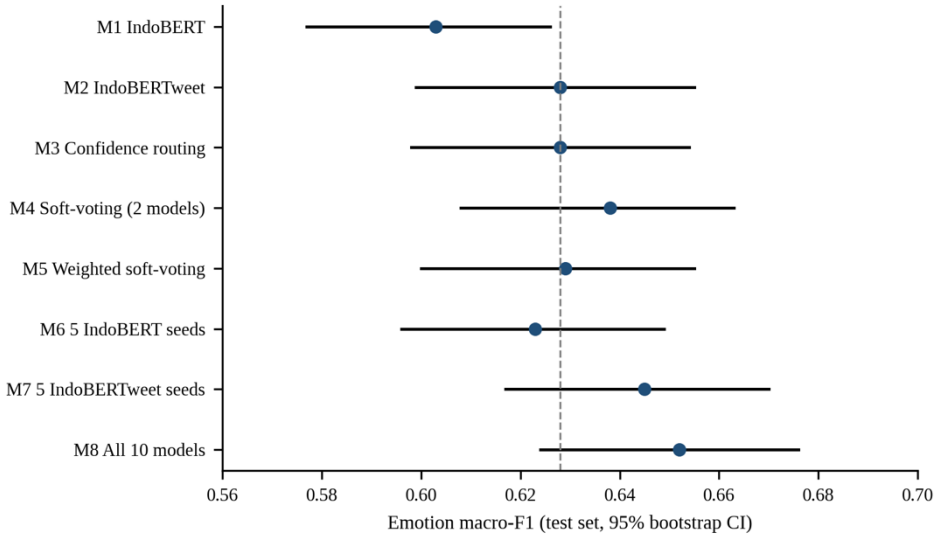


Figure 2 Emotion Macro-F1 of each method with 95% bootstrap intervals

Routing. Confidence fallback routing used IndoBERTweet as the primary model for emotion and IndoBERT for sentiment, with thresholds of 0.65 and 0.90 chosen on the validation set. It sent only 4.9% (emotion) and 7.5% (sentiment) of the test reviews to the secondary model, and its emotion macro-F1 of 0.628 is identical to that of IndoBERTweet alone (difference −0.000, interval −0.008 to 0.006). Routing by task cannot help either, because IndoBERTweet was the better or equal model on both tasks. A first routing prototype evaluated before the pipeline was corrected had shown a nominal gain that was not significant (McNemar p = 0.06), and no gain appeared under the de-duplicated protocol. Throughout this paper, the number of forward passes for confidence fallback routing is reported as the empirically observed average over the test set rather than a worst-case bound, because the average reflects the actual computational cost realized under the measured 4.9% (emotion) and 7.5% (sentiment) fallback rates and remains directly comparable to the fixed pass counts of the other methods in Table 3. Table 5 therefore reports averages of ≈2.12 passes for the dedicated-head configuration and ≈1.05 passes for the derived-sentiment configuration.

 ISSN:2302-8149
e-ISSN:2540-9719

Averaging. Soft-voting of one IndoBERT and one IndoBERTweet model raised emotion macro-F1 to 0.638 (+0.009, interval −0.005 to 0.026), a difference that is not significant, and a weighted variant with a validation-tuned weight added only +0.001. Averaging five IndoBERT seeds did not help (−0.005, interval −0.030 to 0.021), averaging five IndoBERTweet seeds gave +0.017 with a lower bound of 0.000, and averaging all ten models gave 0.652 (+0.024, interval 0.005 to 0.044) with an accuracy of 0.729 against 0.705. A greedy ensemble selected on the validation set from every model trained in the study (XLM-RoBERTa large seeds 42 and 44, IndoBERTweet seed 42, and IndoBERT seed 46) reached 0.658, which is not significantly different from the ten-model ensemble ($p = 0.113$). Selecting one model per input or per task therefore does not move the score. Averaging probabilities shows a less uniform pattern: only the ten-model ensemble produced a confidence interval that excludes zero, while two-model voting (+0.009) and the five-seed IndoBERT ensemble (−0.005) did not reach significance, and the latter numerically decreased the score. The effect of averaging therefore appears to depend on the specific composition of the ensemble (which encoders and how many seeds are combined) rather than on model count alone; because encoder type and seed count were not varied independently in this study, their separate contributions cannot be disentangled here. Two-model voting captures roughly 40% of the ten-model gain at one fifth of the cost, but this gain is not significant on its own.

Register of the text. The premise that IndoBERT should be preferred for formal reviews and IndoBERTweet for informal ones was tested by splitting the test set with the slang-token criterion. Table 4 and Figure 3 show that IndoBERTweet was better on both subsets, but its advantage was concentrated on informal text, where it led IndoBERT by 10.7 macro-F1 points (0.629 versus 0.522) against 2.3 points on formal text (0.643 versus 0.620). On sentiment the two models were level in both subsets. There is therefore no region of the input space in which a router could profitably choose IndoBERT, which explains the null result of the routing strategies, while the size of the informal-text gap is consistent with the social-media pre-training of IndoBERTweet being the property that matters for this domain. Soft-voting was below IndoBERTweet alone on informal text (0.621 versus 0.629) and above it on formal text (0.653 versus 0.643). Because no intervals were computed for the subsets, these differences are descriptive, but they suggest that a selection between IndoBERTweet alone and the two-model vote could be worth testing in future work.

Table 4 Macro-F1 by register of the review (original test set)

Subset	n	Model	Emotion macro-F1	Sentiment macro-F1
Informal	1,739	IndoBERT	0.522	0.770
		IndoBERTweet	0.629	0.770
		Soft-voting	0.621	0.787
		Gap IndoBERTweet - IndoBERT	+10.7 points	≈ 0
Formal	2,475	IndoBERT	0.620	0.801
		IndoBERTweet	0.643	0.802
		Soft-voting	0.653	0.810
		Gap IndoBERTweet - IndoBERT	+2.3 points	≈ 0

 ISSN:2302-8149
e-ISSN:2540-9719

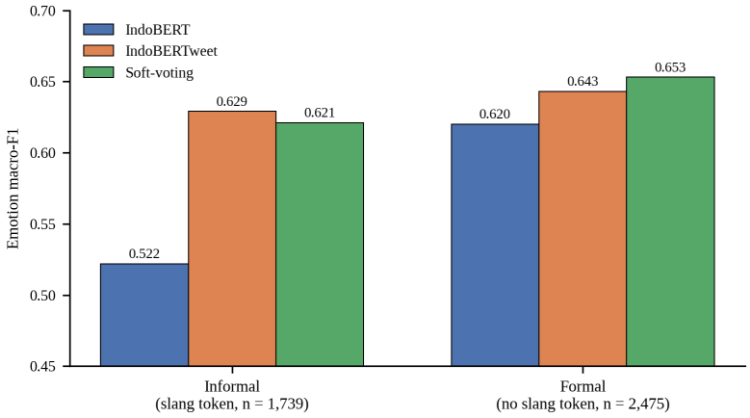


Figure 3 Emotion Macro-F1 by register of the review (original test set)

Sentiment derived from emotion. Table 5 compares, for each configuration, the sentiment obtained from a dedicated head with the sentiment derived from the emotion prediction through Equation (2). The derived sentiment was never worse in any of the seven configurations. For the single IndoBERTweet model it was significantly better (macro-F1 0.807 versus 0.796; McNemar p = 0.012), and for the ten-model ensemble it produced the highest sentiment score of the study, 0.821, with 10 instead of 20 forward passes, although that difference is not significant on its own (p = 0.084). The dedicated heads also contradicted the emotion output on 14.6% of the reviews when IndoBERT predicted sentiment and IndoBERTweet predicted emotion, and on 10.2% when both tasks used soft-voting, producing pairs such as a positive sentiment with an angry emotion. The derived scheme has no contradictions by construction. In practical terms, a deployment needs only the emotion classifier, and the sentiment head can be discarded without a measurable loss on this dataset.

Table 5 Dedicated sentiment head compared to sentiment derived from emotion (original test set, macro-F1)

Configuration	Dedicated head (passes)	Derived from emotion (passes)	Difference	McNemar p
IndoBERT (seed 42)	0.797 (2)	0.798 (1)	+0.001	0.831
IndoBERTweet (seed 42)	0.796 (2)	0.807 (1)	+0.011	0.012
Confidence fallback routing	0.804 (≈2.12)	0.809 (≈1.05)	+0.005	0.366
Soft-voting IndoBERT + IndoBERTweet	0.808 (4)	0.811 (2)	+0.003	0.450
IndoBERT sentiment + IndoBERTweet emotion	0.797 (2)	0.807 (1)	+0.011	0.077
Ensemble of 5 IndoBERTweet seeds	0.811 (10)	0.818 (5)	+0.006	0.143
Ensemble of all 10 models	0.814 (20)	0.821 (10)	+0.007	0.084

Ablations. Table 6 summarises the variants trained on the same split. To keep comparisons controlled, its reference row reports the mean over three of the five main-protocol seeds (42–44) rather than the full five-seed mean of 0.627 ± 0.003. None of them exceeded a macro-F1 of 0.63 for a single model, and the between-seed standard deviations (0.001 to 0.013) are of the same order as the differences that would be needed to call any variant an improvement. A large IndoBERT encoder (three times the parameters) and XLM-RoBERTa large scored 0.628 and 0.623. Minimal pre-processing that preserves emojis and slang changed nothing. Label smoothing lowered the score by almost eight points, and the variant that combines mean–max pooling, focal loss, and a higher head learning rate was significantly worse at the ensemble level: its ten-model ensemble reached 0.637, which is 0.016 below the ten-model ensemble of the main protocol (interval −0.032 to −0.002). A hyperparameter search that had appeared to add 2.3 points on the preliminary split lowered the score

 ISSN:2302-8149
e-ISSN:2540-9719

by 0.5 points when it was re-tested on the corrected pipeline. In the preliminary screening, oversampling, undersampling, and text augmentation on frozen features stayed between 0.599 and 0.616 against 0.610 for the baseline, and merging love into happy gave 0.6496 against 0.6518 for six classes.

Table 6 Ablation study on the emotion task (Macro-F1, Mean ± SD over seeds)

Configuration	Seeds	Macro-F1	Change vs reference (points)	Note
IndoBERTweet, main protocol (reference)	3	0.627 ± 0.002	–	Mean over 3 of the 5 main-protocol seeds (42–44), for comparability with other ablation rows
Minimal pre-processing (emojis, punctuation, slang kept)	3	0.626 ± 0.012	0.0	No effect
Label smoothing 0.1	3	0.550 ± 0.003	−7.7	0.05 and 0.2 also lower
Optuna hyperparameters	3	0.622 ± 0.001	−0.5	Preliminary +2.3 did not hold
Minimal pre-processing + smoothing + Optuna	3	0.569 ± 0.007	−5.8	Dominated by smoothing
IndoBERT large	3	0.628 ± 0.013	0.0	3 times the parameters
XLM-RoBERTa large	3	0.623 ± 0.003	−0.4	Multilingual, no gain
Mean–max pooling + focal loss + higher head LR	5	0.619 ± 0.009	−0.8	10-model ensemble 0.637 (−0.016)
Remove 3,817 flagged training reviews	3	0.630 ± 0.003	+0.3	No effect
Remove top 10% or 25% flagged; down-weight flagged	3	0.62 to 0.63	≈0	No effect
Relabel flagged reviews automatically	3	≈0.60	≈−2.5	Introduces new errors
Greedy ensemble over all trained models	–	0.658	+3.0 vs M2	p = 0.113 vs M8

Label reliability and the performance ceiling. Several indicators converge on the same conclusion regarding the reliability of the labels available for this dataset. Emotion agreement between the original annotators was 0.604, 68 of 346 groups of duplicated texts (19.7%) carried conflicting labels, and confident learning flagged 25.9% of the training reviews as probable label errors. Against this background, the convergence of single-model variants that did not degrade performance at 0.62 to 0.63 macro-F1 is what one would expect if the labels, rather than the architecture, set the limit: larger encoders, multilingual pre-training, tuning, alternative losses, and imbalance techniques left the score unchanged or lower, and only averaging models, which reduces variance, produced a reliable gain. This reading is consistent with the evidence but not proven by it. Removing the flagged training reviews did not help when the models were evaluated on the original

 ISSN:2302-8149
e-ISSN:2540-9719

test labels (0.630 ± 0.003), which may simply reflect that the test labels carry the same annotation conventions, and relabelling flagged reviews automatically lowered the score to about 0.60 and introduced new errors, for example “suka aplikasi ini” (I like this app) relabelled as love. A definitive test would re-annotate the training data with a written guideline and retrain, a direction left for future work.

Adequacy of the label agreement for model comparison. Three observations from the literature indicate that the labels are adequate for the relative comparisons made here, although not for claims about absolute accuracy. First, on the scale of Landis and Koch the emotion kappa of 0.604 lies at the upper edge of the moderate band and the sentiment kappa of 0.625 in the substantial band [10], and the dataset authors describe the average agreement as moderate [9]. Second, this level is typical or better for affective annotation: Bobicev and Sokolova summarise studies with kappa between 0.26 and 0.75 in which classifiers were nevertheless trained and compared [23], and the Gojek sentence-level sentiment annotation reached a Fleiss' kappa of 0.636 on a subset of 4,384 instances [13]. Third, Artstein and Poesio note that no threshold appropriate for computational linguistics has been settled, with 0.67 cited as the accepted reference threshold [22], so a value just below it calls for caution rather than rejection. Agreement also bounds what a classifier can be expected to reach: in the study of Bobicev and Sokolova the classification F-score was 0.794 on the posts on which all annotators agreed exactly and 0.571 on all posts [23]. The single-model scores reported here, an emotion accuracy of 0.69 to 0.71 and a macro-F1 of 0.60 to 0.63, sit close to the 0.65 accuracy and 0.62 F1-score that Cahya et al. reported for IndoBERT on the original, unbalanced version of the same dataset [20]. The labels are therefore treated as usable for ranking models, and the absolute scores are not claimed to carry over to a dataset with higher agreement.

Comparison with earlier reports. Scores reported for IndoBERT on Indonesian app reviews are much higher on polarity tasks, for example 81% accuracy for PLN Mobile [4], 88 to 90% for Shopee and Bukalapak [6], and 97.28% for Mobile JKN [5], but those studies used two or three classes and, in the first two cases, lexicon-generated labels. Emotion studies on other data report accuracies between 0.73 and 0.80 for five or six classes [7], [19]. A score of 0.86 on this dataset after oversampling [20] should not be compared with the results reported here, because the balanced corpus was built before the split. The contribution of this study is therefore not a higher absolute score but a protocol in which the comparisons between encoders and combination strategies carry confidence intervals.

Practical implications. For a developer who needs one emotion classifier, IndoBERTweet alone gives a macro-F1 of 0.628 with a single forward pass, and sentiment can be read from its output at no extra cost, with a derived macro-F1 of 0.807. Two-model soft-voting offers a small, non-significant gain at twice the cost, and averaging ten models offers the only significant gain at ten times the cost. Since no input region favours IndoBERT, storing a second model for routing is not justified by these data.

Limitations. The study uses one dataset whose sentiment labels are defined from emotion, so the derived-sentiment result should not be generalized to datasets that annotate sentiment independently. The informal register is approximated by the presence of a slang-dictionary token, and the subset scores in Table 4 are descriptive because no intervals were computed for them. The strategies that combine one IndoBERT and one IndoBERTweet model were evaluated on a single pair of models rather than on all five seed pairs. Multilingual BERT appeared only in preliminary experiments on an earlier data split, and XLM-RoBERTa large served as the multilingual reference in the final protocol. Finally, label reliability was assessed through agreement statistics and model-based flags rather than through re-annotation of the training data, which was outside the scope of this study.

5 Conclusion

This study compared IndoBERT and IndoBERTweet for emotion and sentiment classification of Indonesian app reviews under a de-duplicated, seed-controlled protocol with confidence intervals. IndoBERTweet was about three macro-F1 points better than IndoBERT on emotion (0.627 versus 0.596 over five seeds; five-seed t-test, p=0.005), with the advantage concentrated on reviews containing slang and small on the rest, while the two models were level on sentiment. Because IndoBERTweet was at least as good on both registers, routing by confidence or by task gave no gain

 ISSN:2302-8149
e-ISSN:2540-9719

(difference 0.000). Averaging the probabilities of ten models raised emotion macro-F1 by 0.024 (interval 0.005 to 0.044), whereas two-model soft-voting gave a non-significant +0.009. Sentiment derived from the predicted emotion was never worse than a dedicated head across seven configurations, was significantly better for a single IndoBERTweet model, was consistent with the emotion label by construction, and halved the inference cost; this result holds for a dataset whose sentiment labels are defined from emotion. Larger encoders, tuning, alternative losses, imbalance techniques, and training-label cleaning all plateaued at 0.62 to 0.63, which is consistent with label reliability, with an original emotion kappa of 0.604, limiting attainable performance. For practitioners, an IndoBERTweet emotion classifier is recommended, with sentiment read from the emotion label, and two-model voting or seed ensembles where the extra inference cost is acceptable. Future work should train and evaluate multilingual BERT under the final protocol, evaluate the combination strategies on all seed pairs, test a selection between IndoBERTweet alone and the two-model vote by register, and re-annotate the training partition with a written guideline to test whether the performance ceiling can be lifted.

References

[1] J. Dąbrowski, E. Letier, A. Perini, and A. Susi, "*Analysing App Reviews for Software Engineering: A Systematic Literature Review*," *Empir. Softw. Eng.*, Vol. 27, No. 2, p. 43, Mar. 2022, DOI: 10.1007/s10664-021-10065-7.

[2] E. V. A. Karyoto, Y. T. Wiranti, and M. I. A. Putera, "Pengaruh *Behavioral Intention* terhadap *Use Behavior* pada Penggunaan Aplikasi Gojek," *Teknika*, Vol. 13, No. 1, pp. 109–119, Mar. 2024, DOI: 10.34148/teknika.v13i1.761.

[3] M. D. Darullah and R. Tanamal, "Faktor yang Paling mempengaruhi Keputusan dalam menggunakan TikTok Shop," *J. Econ. Manag. Account. Technol.*, Vol. 6, No. 2, pp. 166–175, Aug. 2023, DOI: 10.32500/jematech.v6i2.2846.

[4] Y. Asri, D. Kuswardani, W. N. Suliyanti, Y. O. Manullang, and A. R. Ansyari, "*Sentiment Analysis based on Indonesian Language Lexicon and IndoBERT on User Reviews PLN Mobile Application*," *Indones. J. Electr. Eng. Comput. SCI.*, Vol. 38, No. 1, p. 677, Apr. 2025, DOI: 10.11591/ijeecs.v38.i1.pp677-688.

[5] Tarwoto, R. Nugroho, N. Azka, and W. S. R. Graha, "Analisis Sentimen Ulasan Aplikasi Mobile JKN di *Google PlayStore* menggunakan *IndoBERT*," *J. JTIK J. Teknol. Inf. Dan Komun.*, Vol. 9, No. 2, pp. 495–505, Jan. 2025, DOI: 10.35870/jtik.v9i2.3340.

[6] K. C. Pradhisa and R. Fajriyah, "Analisis Sentimen Ulasan Pengguna *E-Commerce* di *Google Play Store* menggunakan Metode *IndoBERT*," *Build. Inform. Technol. SCI. BITS*, Vol. 6, No. 1, Jun. 2024, DOI: 10.47065/bits.v6i1.5247.

[7] A. C. Saputra, A. S. Saragih, and D. Ronaldo, "Prediksi Emosi dalam Teks Bahasa Indonesia menggunakan Model *IndoBERT*," *J. Teknol. Inf. J. Keilmuan dan Apl. Bid. Tek. Inform.*, Vol. 19, No. 1, pp. 1–15, Jan. 2025.

[8] A. S. Rizky and E. Y. Hidayat, "*Emotion Classification in Indonesian Text using IndoBERT*," *Comput Eng Appl J*, Vol. 14, No. 1, pp. 1–11, 2025.

[9] Riccosan and K. E. Saputra, "*Multilabel Multiclass Sentiment and Emotion Dataset from Indonesian Mobile Application Review*," *Data Brief*, Vol. 50, p. 109576, Oct. 2023, DOI: 10.1016/j.dib.2023.109576.

[10] I. Nurjannah and S. M. Siwi, "*Guidelines for Analysis on Measuring Interrater Reliability of Nursing Outcome Classification*," *Int. J. Res. Med. SCI.*, Vol. 5, No. 4, p. 1169, Mar. 2017, DOI: 10.18203/2320-6012.ijrms20171220.

[11] B. Wilie *et al.*, "*IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding*," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, Suzhou, China: Association for Computational Linguistics, 2020, pp. 843–857. DOI: 10.18653/v1/2020.aacl-main.85.

[12] F. Koto, J. H. Lau, and T. Baldwin, "*IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization*," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and

 ISSN:2302-8149
e-ISSN:2540-9719

Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 10660–10668. DOI: 10.18653/v1/2021.emnlp-main.833.

[13] I. N. Aqiilah, R. Saptono, and A. Syaifuddin, "*Sentence-Level Sentiment Analysis of Indonesian App Reviews using IndoBERTweet*," *J. Comput. Theor. Appl.*, Vol. 4, No. 1, pp. 74–91, Jun. 2026, DOI: 10.62411/jcta.16240.

[14] H. Jayadianti, W. Kaswidjanti, A. T. Utomo, S. Saifullah, F. A. Dwiyanto, and R. Drezewski, "*Sentiment Analysis of Indonesian Reviews using Fine-Tuning IndoBERT and R-CNN*," *Ilk. J. Ilm.*, Vol. 14, No. 3, pp. 348–354, Dec. 2022, DOI: 10.33096/ilkom.v14i3.1505.348-354.

[15] C. Hayat, "*Sentiment Analysis of User Perceptions Toward the Gemini Application using the Naïve Bayes Algorithm*," *Sist. J Sist Inf*, Vol. 15, No. 7, pp. 2456–2469, 2026.

[16] S. A. Nugraha, C. Lestari, K. B. Sanjaya, R. A. Naya, and J. Jolie, "*Comparative Analysis of IndoBERT and mBERT for Online Gambling Comment Detection in Indonesian Social Media*," *J. Tek. Inform. Jutif*, Vol. 7, No. 2, pp. 1931–1943, Apr. 2026, DOI: 10.52436/1.jutif.2026.7.2.5677.

[17] Riccosan and K. E. Saputra, "*Multilabel Classification Sentiment Analysis on Indonesian Mobile App Reviews*," *IAES Int. J. Artif. Intell. IJ-AI*, Vol. 14, No. 5, pp. 4226–4234, Oct. 2025, DOI: 10.11591/ijai.v14.i5.pp4226-4234.

[18] G. Z. Nabiilah, I. N. Alam, E. S. Purwanto, and M. F. Hidayat, "*Indonesian Multilabel Classification using IndoBERT Embedding and MBERT Classification*," *Int. J. Electr. Comput. Eng. IJECE*, Vol. 14, No. 1, p. 1071, Feb. 2024, DOI: 10.11591/ijece.v14i1.pp1071-1078.

[19] W. Christian, D. Adamlu, A. Yu, and D. Suhartono, "*Leveraging IndoBERT and DistilBERT for Indonesian Emotion Classification in e-Commerce Reviews*," *Procedia Comput. SCI.*, Vol. 269, pp. 321–330, 2025, DOI: 10.1016/j.procs.2025.08.284.

[20] L. D. Cahya, A. Luthfiarta, J. I. T. Krisna, S. Winarno, and A. Nugraha, "*Improving Multi-Label Classification Performance on Imbalanced Datasets Through SMOTE Technique and Data Augmentation using IndoBERT Model*," *J. Nas. Teknol. dan Sist. Inf.*, Vol. 9, No. 3, pp. 290–298, Jan. 2024, DOI: 10.25077/TEKNOSI.v9i3.2023.290-298.

[21] W. Tang, J. Hu, H. Zhang, P. Wu, and H. He, "*Kappa Coefficient: A Popular Measure of Rater Agreement*," *Shanghai Arch. Psychiatry*, Vol. 27, No. 1, pp. 62–67, 2015, DOI: 10.11919/j.issn.1002-0829.215010.

[22] R. Artstein and M. Poesio, "*Inter-Coder Agreement for Computational Linguistics*," *Comput. Linguist.*, Vol. 34, No. 4, pp. 555–596, Dec. 2008, DOI: 10.1162/coli.07-034-R2.

[23] V. Bobicev and M. Sokolova, "*Inter-Annotator Agreement in Sentiment Analysis: Machine Learning Perspective*," in *RANLP 2017 - Recent Advances in Natural Language Processing Meet Deep Learning*, Incoma Ltd. Shoumen, Bulgaria, Nov. 2017, pp. 97–102. DOI: 10.26615/978-954-452-049-6_015.

[24] C. G. Northcutt, L. Jiang, and I. L. Chuang, "*Confident Learning: Estimating Uncertainty in Dataset Labels*," Aug. 22, 2022, *arXiv*: arXiv:1911.00068. DOI: 10.48550/arXiv.1911.00068.

[25] T. G. Dietterich, "*Ensemble Methods in Machine Learning*," in *Multiple Classifier Systems*, Lecture Notes in Computer Science, Vol. 1857, Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, pp. 1–15. DOI: 10.1007/3-540-45014-9_1.

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "*BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*," May 24, 2019, *arXiv*: arXiv:1810.04805. DOI: 10.48550/arXiv.1810.04805.

[27] R. Dror, G. Baumer, S. Shlomov, and R. Reichart, "*The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing*," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 1383–1392. DOI: 10.18653/v1/P18-1128.

[28] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "*Optuna: A Next-Generation Hyperparameter Optimization Framework*," Jul. 25, 2019, *arXiv*: arXiv:1907.10902. DOI: 10.48550/arXiv.1907.10902.

[29] A. Conneau *et al.*, "*Unsupervised Cross-Lingual Representation Learning at Scale*," Apr. 08, 2020, *arXiv*: arXiv:1911.02116. DOI: 10.48550/arXiv.1911.02116.

ISSN:2302-8149
e-ISSN:2540-9719

[30] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "*Focal Loss for Dense Objec*t *Detection*," Feb. 07, 2018, *arXiv*: arXiv:1708.02002. DOI: 10.48550/arXiv.1708.02002.