Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15472
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Comparison of IndoBERT and SVM Performance in Sentiment Analysis of Digital Education Platforms

Aldina Bonaria Siva[1)], **Robet**[2)], **Leony Hoki**[3)]
[1,2,3)] Department of Informatics, STMIK Time, Medan, Indonesia
[1)]sembiringaldina@gmail.com, [2)*]robertdetime@gmail.com, [3)]leony.hoki@gmail.com

Submitted :Oct 22, 2025 | **Accepted** : Dec 5, 2025 | **Published** : Jan 02, 2026

Abstract: Sentiment analysis on user-generated reviews is essential for understanding the quality and effectiveness of digital education platforms. This study compares the performance of Support Vector Machine (SVM) and IndoBERT in classifying sentiments from Ruangguru user reviews. The original dataset contains 111,838 reviews, from which a stratified sample of 10,000 entries was selected for experimentation to maintain class proportion. Text preprocessing applied standard/light normalization (case folding and light cleaning, handling URLs/users/hashtags and repetition) without stopword removal to preserve polarity cues. Auto labels are validated on 139 manually annotated samples (accuracy 0.763, Cohen's κ 0.644), indicating reliable yet imperfect alignment. To ensure a fair, leakage-safe comparison, we use a fixed 20% standard test split for all models; within the remaining data, 10% is used for validation, and IndoBERT checkpoints are selected based on validation macro-F1 (early stopping). The SVM baseline combines word- and character-level TF-IDF with class-balanced LinearSVC and grid search, achieving accuracy 0.888 and macro-F1 0.543, strong on positives but limited for the neutral class. IndoBERT yields more balanced performance: the class-weighted variant attains the best macro-F1 0.601 (accuracy 0.857), while the baseline reaches the highest IndoBERT accuracy (0.867) with macro-F1 0.596. These results show that Transformer models provide a more balanced trade-off under severe imbalance, whereas SVM remains a competitive accuracy-oriented baseline. In practice, platforms should prioritize macro-F1, use optimized IndoBERT when minority opinions matter, and invest in expanded manual labeling and advanced imbalance handling to improve neutral detection further.

Keywords: IndoBERT; Ruangguru; Sentiment Analysis; SVM; TF-IDF

INTRODUCTION

The advancement of information and communication technology has significantly accelerated the adoption of digital learning platforms, reshaping how students in Indonesia access educational resources (Anam et al., 2023). Services such as Ruangguru have become central to online tutoring ecosystems, offering flexible, scalable, and interactive learning environments that gained massive momentum during and after the COVID-19 pandemic (Mushtaha et al., 2022).

As user engagement increases, platforms accumulate thousands of text-based reviews that capture perceptions, expectations, and criticisms, which are essential for evaluating service quality and guiding strategic decisions (Sarasvananda et al., 2022). Extracting actionable insights from such unstructured reviews requires robust sentiment analysis techniques that can handle Indonesian linguistic features, including informal expressions, code-switching, and implicit sentiment cues.

Sentiment analysis, as part of natural language processing (NLP), has become a vital approach for understanding users' opinions expressed in text, enabling automatic classification into positive, neutral, or negative categories that support evidence-based product improvement (Rafiandi Andhika et al., 2025). Traditional machine learning methods, notably Support Vector Machine (SVM) and Naïve Bayes, remain widely used in Indonesian sentiment analysis due to their stability, interpretability, and strong performance when combined with TF-IDF features and class-weight strategies in imbalanced datasets (Jannah & Kusnawi, 2024). However, these classical

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

models often struggle to capture deeper semantic and contextual meaning in Indonesian sentences, limiting their robustness when reviews contain ambiguous sentiment cues or informal linguistic patterns.

Recent advancements in deep learning have introduced transformer-based architectures such as Bidirectional Encoder Representations from Transformers (BERT). IndoBERT, a variant pre-trained specifically on large-scale Indonesian corpora, delivers richer contextual embeddings and has demonstrated superior performance in various Indonesian NLP tasks, including educational sentiment analysis and public policy evaluation (Koto et al., 2020). Despite these promising results, direct comparisons between IndoBERT and classical machine learning models on real-world Indonesian digital education reviews remain scarce, particularly in settings with noisy user-generated text, automatically assigned sentiment labels, and skewed class distributions.

To address this gap, this study conducts a controlled empirical comparison between IndoBERT and SVM on Ruangguru user reviews using a leakage-safe experimental protocol. From 111,838 reviews, we draw a stratified sample of 10,000 entries that preserves class proportions. We performed text preprocessing and applied light normalization (case folding, handling URLs/users/hashtags, limiting repeated characters) without stopword removal. Sentiment labels are automatically derived from rating scores and validated on 139 manually annotated samples (accuracy = 0.763; Cohen's K = 0.644), indicating reliable yet imperfect alignment. All models are evaluated on a fixed 20% standard test split; within the remaining data, 10% is used for validation, and IndoBERT checkpoints are selected by validation macro-F1 (early stopping). SVM uses word and character-level TF-IDF with class-balanced LinearSVC and stratified 5-fold grid search. We report accuracy, precision, recall, and macro-F1 to reflect performance under severe imbalance.

This study provides a controlled, leakage-safe baseline for sentiment classification on Indonesian digital learning platforms by systematically comparing IndoBERT and SVM on a standard dataset using a unified protocol (light normalization without stopword removal, a fixed 20% standard test split, 10% validation, and early stopping by validation macro-F1). While prior work has applied SVM (Fitri et al., 2020) and shown strong results with IndoBERT in Indonesian sentiment tasks (Octarini et al., 2025) in isolation, to our knowledge, this is among the first controlled comparisons that incorporate class-weighting/sampling strategies under realistic, user-generated educational reviews. The findings offer practical guidance for researchers, developers, and policymakers to strengthen sentiment-driven monitoring and quality assurance in Indonesia's digital education ecosystem.

LITERATURE REVIEW

Research on Indonesian sentiment analysis in recent years has highlighted two main methodological approaches: the classical feature-based approach (TF-IDF, Word2Vec combined with ML algorithms such as SVM, Naïve Bayes, or Logistic Regression) and the modern contextual representation-based approach using pre-trained Transformer models, such as IndoBERT.

Several studies show that the effectiveness of classical methods remains quite high on relatively clean datasets, with dominant label distributions, and in simple linguistic contexts; for example, (Suharman et al., 2025) reported that SVM with TF-IDF can compete with modern approaches in the domain of mobile banking applications. These findings confirm that classical methods remain relevant in limited domains, but are less flexible when dealing with informal text or implicit contexts.

Although Transformer-based architectures such as BERT are renowned for capturing semantic nuance and long-range contextual dependencies, empirical studies show that classical machine-learning methods can remain highly competitive under certain data conditions. The study (Ulinuha et al., 2025) demonstrates that SVM and Naïve Bayes can match or even surpass BERT on small or imbalanced datasets, indicating that model performance is shaped not only by architectural sophistication but also by the interaction between text representation, data characteristics, and class-balancing strategies.

Hybrid research has emerged to address the complexity of heterogeneous datasets by combining contextual embeddings with deep learning architectures such as LSTMs or TCNs. (Octarini et al., 2025) reported that this combination is capable of handling variations in topics and language styles, but is not always superior to classical methods with balancing techniques on small datasets.

Supporting these findings, (Nur et al., 2025) report that the classical method combined with SMOTE remains competitive in the social/environmental domain when the dataset is limited and the class distribution is imbalanced. A synthesis of these findings shows that a model's success is highly dependent on data characteristics, dataset size, and balancing strategies, rather than solely on the architecture's or model type's complexity.

Several studies show that automatic/lexicon-based labeling is often used due to its ease and its suitability for large-scale data. Still, results indicate a tendency toward class distribution bias, particularly the dominance of neutral classes, raising serious questions about the validity of labels without manual validation. This underlies the need for label validation procedures, primarily when datasets are used for training ML or Transformer models (Imtihan et al., 2025).

The handling of unbalanced data remains minimal; most research continues to focus on the two dominant classes, while minority classes are rarely analyzed in depth (Munthe & Levianti, 2025). Research that integrates

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15472
e-ISSN : 2541-2019
p-ISSN : 2541-044X

contextual representation, balancing techniques, and label validation simultaneously is still scarce. Documentation of errors or mismatched cases is also rarely done, so the reliability of results in real datasets cannot be guaranteed in general.

Review/opinion datasets in Indonesia are often skewed; for example, only 5–10% of hotel reviews are neutral/negative, so models such as IndoBERT tend to fail to capture minority classes (Singgalen, 2025). Several studies have attempted to address imbalance with balancing technologies such as SMOTE or SMOTE-ENN, for example (Nawulansih et al., 2025), who conducted sentiment analysis of DANA app reviews on Google Play Store using SMOTE to handle data imbalance before classification with SVM.

From this literature synthesis, the research gaps are apparent: automatic label validation remains minimal, minority/neutral class analysis is rarely conducted in depth, the integration of contextual representation with balancing techniques and data validation is seldom tested simultaneously, error documentation is almost non-existent, and generalization across domains remains limited.

This literature synthesis serves as the basis for designing experiments that compare SVM and IndoBERT on the same dataset, using stratified sampling, class-weight balancing, manual label validation, and per-class error analysis. This approach not only compares algorithm performance but also builds a transparent, credible, and relevant analytical framework for real-world data conditions.

Table 1. Comparative Analysis of Previous Studies and Research Gap Identification

Study	Domain	Method	Limitation	Gap Filled (this work)
(Suharman et al., 2025)	Mobile banking (ID)	SVM + TF-IDF	Relatively clean data; not tested on noisy UGC	EdTech UGC with noisy text and extreme imbalance
(Ulinuha et al., 2025)	Multi-domain (IMDb, Amazon, Gojek)	BERT variants vs SVM/NB	Single-run evaluation; no CV; no anti-leakage; no label validation; limited per-class analysis	Stratified splits, fixed 20% test and 10% validation, macro-F1 early stopping, and fair baselines
(Octarini et al., 2025)	Hybrid (BERT + LSTM/TCN)	Hybrid DL	Not consistently superior; no weak-label validation	Auto-label validation using a manual subset
(Nur et al., 2025)	Social/environmental	SVM + SMOTE	Classical-focused; no Transformer comparison	Direct comparison between SVM and IndoBERT on the same dataset
(Nawulansih et al., 2025)	App reviews (DANA)	SVM + SMOTE/ENN	No per-class evaluation & no anti-leakage	Macro-F1 and minority-class reporting with safe evaluation protocol
Ours	EdTech platform reviews (Indonesia) — UGC, noisy, imbalanced	Auto-label + 139 manual validation; SVM TF-IDF; IndoBERT with balancing	Limited auto-labels, single-domain data, fixed hyperparameters, and a single test split	Auto-label validation, anti-leakage evaluation (fixed 20% test), per-class analysis, imbalance handling, and fair SVM–IndoBERT comparison

METHOD

Research Stages

This study employs a quantitative, comparative experimental approach to evaluate the performance of two machine learning algorithms: Support Vector Machine (SVM) and IndoBERT (Kono et al., 2025). User reviews from digital education platforms are classified into three sentiment categories: positive, neutral, and negative. To ensure reliable and unbiased evaluation across classes, the study employs stratified sampling, class-balancing, and manual label validation.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15472
e-ISSN : 2541-2019
p-ISSN : 2541-044X

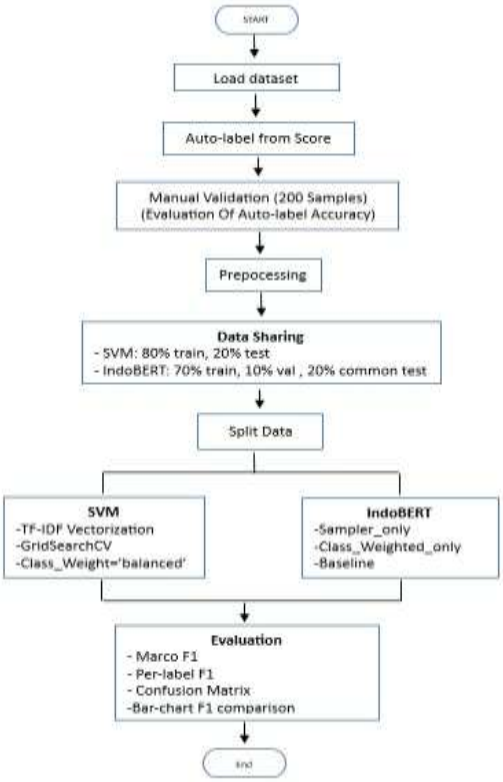


Fig. 1 Research Stages

Load Dataset

The dataset consists of user reviews from the Ruangguru digital education platform. Each entry contains textual review content and a numerical rating score. The data were loaded using the Pandas library in Python, ensuring that both columns were available for further processing. An initial exploration was conducted to understand the score distribution and detect class imbalance. Numerical scores were mapped into three sentiment categories, positive, neutral, and negative, following standard practices in prior studies that use star ratings as proxy sentiment labels through weak supervision (Ashbaugh & Zhang, 2024).

This approach enables automatic processing of large datasets while providing a basis for manual validation to maintain label accuracy, particularly in ambiguous or minority-class cases.

Automatic Labeling from Rating Score

Each review is automatically labeled based on its rating score: scores ≥ 4 are labeled positive, scores = 3 are labeled neutral, and scores ≤ 2 are labeled negative (Masitoh et al., 2025). Since these labels are derived from rating numbers rather than semantic understanding, they are further refined through manual validation. This rule serves only as weak supervision, not as an absolute truth, so its validity is tested quantitatively and qualitatively.

Manual Validation (200 Samples) for Auto-Label Accuracy

We drew a stratified random sample of user reviews for manual validation; after quality control, 139 reviews remained and were used to assess the rating-to-sentiment mapping (accuracy 0.763, Cohen's κ 0.644). The Kappa value was calculated using the following formula:

$$k = \frac{p_0 - p_e}{1 - p_e} \quad (1)$$

where p_o is the proportion of actual agreements and p_e is the proportion of random agreements. The confusion matrix for automatic versus manual labeling is formulated as:

$$CM = \begin{bmatrix} TP_{neg} & FN_{neg\to new} & FN_{neg\to pos} \\ FP_{neu\to neg} & TP_{neu} & FN_{neu\to pos} \\ FP_{pos\to neg} & FP_{pos\to neu} & TP_{pos} \end{bmatrix}. \quad (2)$$

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15472
e-ISSN : 2541-2019
p-ISSN : 2541-044X

In addition to quantitative metrics, this study includes mismatched cases that show concrete examples where automatic and manual labels differ. This analysis was conducted to assess semantic ambiguity and explain the causes of labeling errors, particularly in the neutral class.

Preprocessing Teks

The review text is processed through the stages of case folding and light cleaning, handling URLs/users/hashtags, and removing repetition without stopword removal to preserve polarity cues. For the SVM model, the preprocessed text is converted to a TF-IDF representation. For the IndoBERT model, the text is reprocessed using the WordPiece tokenizer. The TF-IDF formula used follows the general equation:

$$TF - IDF(t,d) = tf(t,d).log\left(\frac{N}{df(t)}\right) \quad (3)$$

with $tf(t,d)$is the frequency of occurrence of the term t on the document d, N total number of documents, and $df(t)$ number of documents containing that term.

Data Sharing

Data splitting. We perform a two-stage split. First, a fixed 20% standard test split is held out from the outset and used jointly by SVM and IndoBERT. Second, from the remaining 80%, we reserve 10% as a stratified validation set for IndoBERT; SVM performs model selection via stratified 5-fold cross-validation on the training portion only. In imbalance handling, we are given a severe class imbalance (notably in the neutral class). We evaluate three IndoBERT training configurations: (i) baseline (no reweighting/sampling), (ii) class-weighted loss (Cross-Entropy with class weights and a mild neutral boost), and (iii) sampler-only (no class weights; WeightedRandomSampler on the training set). For SVM, we use LinearSVC with class_weight='balanced' and no sampling.

Weighted sampling. In the sampler-only configuration, the selection probability of each training instance is set proportional to the inverse of its class frequency in the training set. Let y_i be the class of the sample $i, n_{y_i}$the number of training samples in class y_i, and M is the size of the training set. We define unnormalized per-sample weights and the resulting sampling probabilities as

$$w_i \alpha \frac{1}{ny_i}, \quad p_i = \frac{w_i}{\sum_{k=1}^{M} w_k}. \quad (4)$$

This strategy increases the chance that minority examples appear in mini-batches without creating synthetic data.

Class-Weighted Loss. Independent of sampling, the model can also be trained with class weights in the loss. Let N denotes the number of training samples, K the number of classes, and n_c the number of training samples in class c. We compute the base class weights as

$$w_c = \frac{N}{K.n_c}. \quad (5)$$

To emphasize the hardest class (neutral), we apply a mild neutral boost β (≈ 1.3) to obtain the final weights.

$$w_c = \begin{cases} \beta.\dfrac{N}{K.n_c} \\ \dfrac{N}{K.n_c} \end{cases}, \quad c = neutral, otherwise.$$

Weighted Cross-Entropy. Training minimizes weighted cross-entropy over a mini-batch of size B:

$$L = -\frac{1}{B}\sum_{i=1}^{B}\sum_{c=1}^{k} w_c.y_{ic} \log p_{ic} \quad (6)$$

Where y_{ic} is the one-hot target for the sample i and class c, and p_{ic} is the model's softmax probability.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15472
e-ISSN : 2541-2019
p-ISSN : 2541-044X

SVM and IndoBERT Model Training

SVM: The baseline model was built using LinearSVC with features derived from combining TF-IDF word-level and char-level features via FeatureUnion. Parameter determination was performed using GridSearchCV with parameter value search C ∈ {0.25,0.5,1,2}. The objective function of grid search is written as:

$$C^* = \arg\max_C F1_{macro}^{(CV)}. \quad (7)$$

Training using class_weight='balanced' to reduce class bias and evaluation performed on the fixed 20% standard test split.

IndoBERT: IndoBERT-base is used as the main transformer-based model. Tokens are processed using the WordPiece tokenizer, and training is performed using AdamW and a linear warmup/decay scheduler. In the context of NLP theory, IndoBERT's behavior is explained through a self-attention mechanism formulated as:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (8)$$

which enables the model better to understand context and inter-word relationships than frequency-based models. Training was performed with early stopping based on the macro-F1 score on the validation set. Three balancing configurations (sampler-only, class-weight-only, and baseline) were tested to assess the impact of the balancing mechanism on the performance of the neutral class.

Model Evaluation and Results Analysis

The evaluation was performed using accuracy, F1 per class, macro-F1, and a confusion matrix. Macro-F1 was calculated as:

$$F1_{macro} = \frac{1}{K}\sum_{c=1}^{k}\frac{2.P_c.R_c}{P_c+R_c}, \quad (9)$$

with P_c and R_c the precision and recall for each class c. The results of SVM and IndoBERT are directly compared on a standard test set to ensure a fair performance comparison. For IndoBERT, the loss curve graph is reported to show the training dynamics and convergence stability. Special attention is given to the neutral class, which is considered the most problematic due to its small size and ambiguous semantics. Error analysis includes identifying misclassification patterns and providing concrete examples of frequently misclassified reviews.

Threats to Validity

This study identifies and discusses four types of validity threats. Internal validity is maintained by using a standard test set to prevent data leakage during training and evaluation. External validity is limited because the dataset comes from the Indonesian education domain, so the research results need to be retested to ensure generalizability to other domains or contexts. Construct validity is strengthened through manual validation of labels derived from rating scores to ensure they accurately reflect actual sentiment. Meanwhile, statistical conclusion validity is maintained through cross-validation, data stratification, comprehensive reporting of evaluation metrics, and consistent balancing techniques to address class imbalance.

Dataset Bias Mitigation

This study recognizes a positive bias in the data distribution. Therefore, sampler, class-weight, and neutral boost approaches are used to balance model learning. The study also presents an analysis showing that residual bias can persist due to the polite language style typical of Indonesian reviews, which can mislead the classifier into classifying as positive. This explanation is written explicitly so that the study meets scientific transparency and accountability standards.

RESULT

This section presents the experimental findings obtained from the sentiment analysis of digital education platform reviews using Support Vector Machine (SVM) and IndoBERT. The results include data validation, model performance metrics, confusion matrices, and training curves. All results reported here represent the output directly generated from the procedures described in the Methods section, without additional interpretation.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15472
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Validation of Automatic Labeling

Evaluation of 139 samples showed that the automatic labeling system demonstrated relatively high agreement with manual labeling, with an accuracy of 0.763 and a Cohen's Kappa of 0.644, indicating substantial agreement.

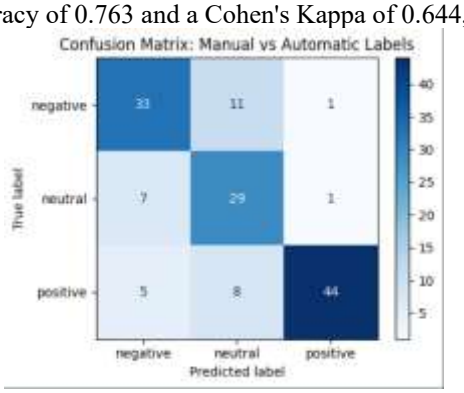


Fig. 2 Confusion Matrix: Manual vs Automatic Labels

The confusion matrix in Figure 2 shows that the best performance is achieved for the positive class (44 correct predictions out of 57 samples). In contrast, most errors occur in shifts between semantically close classes, especially between negative–neutral and positive–neutral. This pattern is natural in sentiment analysis because the neutral class has blurrier boundaries and often overlaps with weak sentiment expressions; in addition, the smaller number of neutral samples likely makes the feature representation for that class less robust, making it easier for the model to misclassify it. Overall, these results show that, despite misalignments in borderline cases, the automated system captures sentiment patterns consistent with human judgments and delivers reliable performance.

SVM Model Performance

The SVM model optimized via 5-fold cross-validation selected the best parameter C = 0.25, yielding a macro-F1 of 0.543, suggesting a challenge in balancing performance across classes. On the standard test set, the model achieved high accuracy (0.888) mainly due to its excellent performance on the positive class (F1 = 0.947), which is the largest class in the data. In contrast, performance for the negative class was moderate (F1 = 0.567), whereas it was deficient for the neutral class (F1 = 0.117).

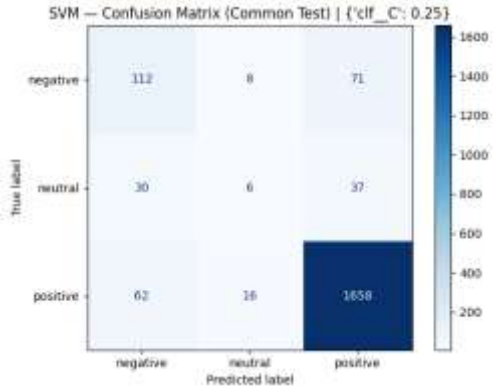


Fig. 3 SVM - Confusion Matrix

The confusion matrix in Figure 3 shows that many negative and neutral samples are misclassified as positive, leading to a high weighted average, but the macro average remains low. This pattern occurs due to extreme class imbalance, which makes it easier for the model to learn patterns of the positive class; the low C value (0.25) also encourages wider margins that tend to ignore minority classes. In addition, the neutral class is linguistically more ambiguous and has far fewer samples, resulting in weaker representation and frequent misclassification by the model. Thus, although SVM appears highly accurate overall, its performance is uneven, particularly on minor and ambiguous categories.

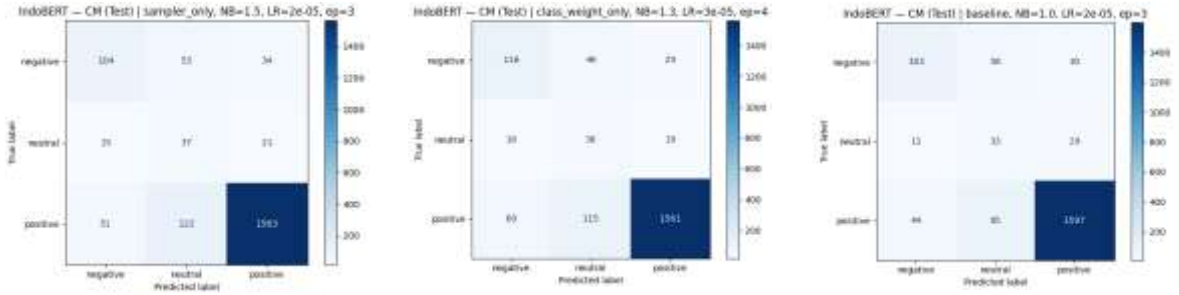
IndoBERT Performance

Three configurations, IndoBERT Sampler Only, Class-Weight Only, and Baseline, were evaluated to assess the effectiveness of data balancing strategies in three-class sentiment classification tasks. All models were trained

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15472
e-ISSN : 2541-2019
p-ISSN : 2541-044X

for a maximum of 10 epochs using early stopping (patience = 2) to prevent overfitting. The evaluation results and confusion matrix of the three models are shown together in Figure 4 to facilitate comparison of error patterns.



IndoBERT Sampler Only IndoBERT Class Weight IndoBERT Baseline

Fig. 4 IndoBERT – CM (Test)

Figure 4. Confusion matrices on the standard 20% test split for IndoBERT under three training configurations: (a) sampler-only (NB=1.5, LR=2e-5, three epochs), (b) class-weight-only (NB=1.3, LR=3e-5, four epochs), and (c) baseline (NB=1.0, LR=2e-5, three epochs). Numbers are counts of test examples. The sampler raises neutral recall but at the cost of precision (many positive-to-neutral confusions). Class-weighting yields the most balanced trade-off across classes, while the baseline preserves the highest positive recall.
(a) Sampler-only: TP = neg 104, neu 37, pos 1563. Recall: neg 0.544, neu 0.507 (highest), pos 0.900. Significant positive→neutral shift (122 cases) improves neutral coverage but depresses precision and macro-F1.
(b) Class-weight: TP = neg 116, neu 36, pos 1561. Recall: neg 0.607 (highest), neu 0.493, pos 0.899. Fewer false-positive neutral than (a), delivering the best macro-F1 among IndoBERT variants.
(c) Baseline: TP = neg 103, neu 33, pos 1597. Recall: neg 0.539, neu 0.452, pos 0.920 (highest). Dominant errors are negative/neutral to positive, explaining the highest IndoBERT accuracy but slightly weaker class balance than (b).
On the standard 20% test split, the three IndoBERT configurations exhibit distinct error profiles. The sampler-only setting increases neutral coverage, as evidenced by the highest neutral recall (0.507), but introduces many positive-to-neutral confusions (122 cases), lowering precision and macro-F1. The class-weight model reduces false-positive neutral predictions and raises negative recall to 0.607, yielding the best macro-F1 among IndoBERT variants. The baseline retains the highest positive recall (0.920) and the highest IndoBERT accuracy, yet still shows a systematic drift from negative/neutral to positive, which limits neutral F1. Collectively, these matrices illustrate the trade-off between minority recall and overall precision, and motivate mild class weighting as the most balanced choice.

Comparative Evaluation

The results in Table 2 show a consistent performance pattern: the positive class always has the highest F1-score, while the neutral class remains the most difficult to classify. The neutral class performs poorly due to the limited number of samples and the ambiguity of linguistic features, which often overlap with weak positive or negative expressions, making it difficult for the model to accurately distinguish them.

Table 2. Summary of SVM and IndoBERT Performance (Test Set)

Model	Accuracy	Macro F1	F1 Neg	F1 Neu	F1 Pos
SVM (best grid C=0.25)	0.888	0.543	0.567	0.117	0.947
IndoBERT – Sampler Only	0.852	0.589	0.576	0.260	0.932
IndoBERT – Class Weight	0.857	0.601	0.603	0.267	0.933
IndoBERT – Baseline	0.867	0.596	0.590	0.255	0.942

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15472
e-ISSN : 2541-2019
p-ISSN : 2541-044X

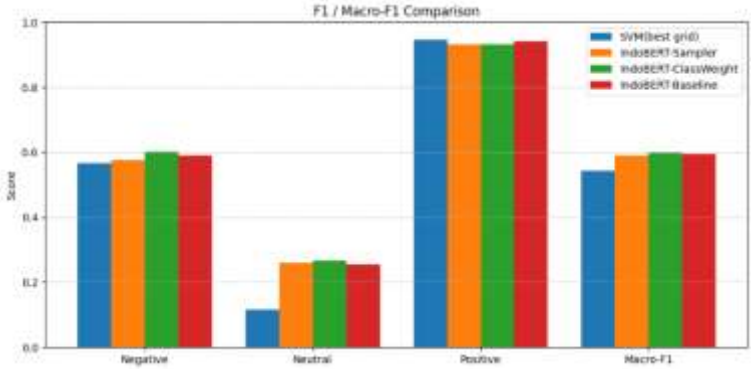


Fig. 5 F1/Macro-F1 Comparasion

SVM (C = 0.25) reaches an accuracy of 0.888 but a lower macro-F1 of 0.543, with F1-neutral of 0.117, indicating strong majority-class bias even with class_weight='balanced'.

For IndoBERT, all configurations are more balanced across classes. Sampler-only increases minority exposure and raises neutral recall, but precision drops; the final scores are F1 = 0.932 (positive), 0.576 (negative), 0.260 (neutral) with macro-F1 0.589, showing that oversampling alone does not sufficiently correct the bias. Class-weight makes gradients more sensitive to minority errors and yields the best macro-F1 (0.601) with F1-neutral 0.267 and F1-negative 0.603. Interestingly, the baseline without explicit balancing remains strong: macro-F1 0.596 (and the highest IndoBERT accuracy), with F1 = 0.942/0.590/0.255 for positive/negative/neutral, very close to the class-weighted variant.

Overall, performance differences are driven primarily by the balancing strategy, and all models share a consistent weakness in the neutral class.

IndoBERT Training Curves

Figure 6 shows the training and validation loss curves for all IndoBERT configurations, illustrating the model's performance over epochs and the epochs at which early stopping was applied. These curves provide essential insights into the model's learning dynamics and stability.

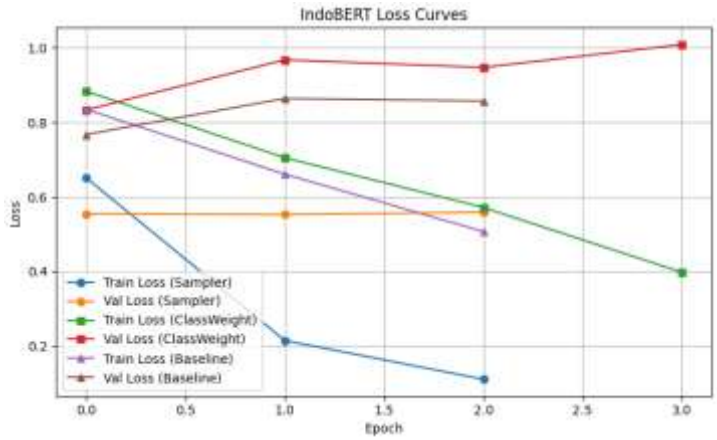


Fig. 6 IndoBERT Loss Curves

Sampler-only shows a rapid drop in training loss, while the validation loss remains almost flat after the first epoch. This indicates limited generalization: oversampling duplicates minority instances, accelerating memorization without yielding steady gains on the validation set.

Class-weight exhibits the most stable dynamics: both training and validation losses decrease across epochs, suggesting that reweighting minority errors strengthens learning signals without distorting the mini-batch distribution. This aligns with its best macro-F1 among the IndoBERT variants.

Baseline displays an unusual trend: training loss increases over epochs, while validation loss rises slightly and then plateaus. This likely reflects optimization/scale effects under severe imbalance, fitting majority patterns while remaining uncertain about minority examples, consistent with strong accuracy but weaker balance on neutral.

Class-imbalance strategies shape not only final scores but also learning dynamics. Class weighting yields the most synchronized train and validation behavior (best generalization), whereas oversampling and the baseline are

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

more prone to stagnation/overfitting, with the sampler (memorization) rapidly and the baseline more gradually or imbalanced.

DISCUSSIONS

The experiment's results, based on the fixed 20% standard test split, show a clear trade-off between overall accuracy and class balance. SVM achieves the highest accuracy (0.888) but the lowest macro-F1 (0.543), with F1-neutral = 0.117. This pattern indicates a strong bias toward the majority positive class despite class-weighted training. The feature space provided by word+character TF-IDF is robust for lexically explicit sentiment (e.g., intensifiers, exclamation patterns). It yields an excellent F1-positive = 0.947, yet it struggles to carve out a reliable decision boundary for neutral, where lexical cues are sparse or ambiguous.

By contrast, IndoBERT variants deliver more balanced performance across classes. The baseline (LR = 2e-5, 3 epochs, early stopping by validation macro-F1) achieves the highest IndoBERT accuracy (0.867) with macro-F1 = 0.596 and F1-neutral = 0.255, suggesting that pretrained contextual representations already capture much of the polarity signal without explicit reweighting. The sampler-only configuration increases recall for neutral (0.507) by amplifying minority exposure within mini-batches, but its precision drops, resulting in macro-F1 = 0.589 and accuracy = 0.852. In other words, oversampling improves coverage of minority patterns but induces noisier decision thresholds. The class-weighted configuration (neutral boost = 1.3; LR = 3e-5; 4 epochs) provides the best overall balance (macro-F1 = 0.601, F1-neutral = 0.267, accuracy = 0.857), showing that mild loss reweighting strengthens minority gradients without distorting mini-batch composition.

Why is neutral persistently difficult? First, linguistic ambiguity is intrinsic: neutral reviews often mix factual descriptions with faintly positive/negative shades ("not bad", "just normal", "good, but…"). Second, data scarcity is substantial—neutral instances are rare in the 10k stratified sample—limiting the model's opportunity to learn diverse neutral prototypes. Third, weak labels (rating→sentiment) introduce label noise: the 139-sample manual check yields accuracy = 0.763 and $\kappa = 0.644$, i.e., good but imperfect alignment. Neutral is the class most prone to mismatch (e.g., 5-star with complaints or 3-star with clearly positive content), and misalignment here disproportionately depresses F1-neutral. Finally, pretrained models can inherit domain priors (e.g., stronger polarity priors than neutrality), which further skews decision boundaries when neutral is under-represented.

Training dynamics clarify the role of each balancing strategy. Sampler-only materially increases minority exposure but can inflate false positives for neutral, lowering precision. Class-weighted loss is gentler: it modifies gradient magnitudes while preserving the natural sample mix, keeping positive/negative precision more stable and lifting neutral recall enough to improve macro-F1. The baseline benefits from strong priors in IndoBERT and clean early stopping; its relatively high IndoBERT accuracy (0.867) suggests that when the goal is overall correctness on imbalanced streams, pretrained representations already go a long way, yet they still leave a gap on neutrality compared to the class-weighted variant.

Methodologically, several choices strengthened validity. First, a leakage-safe protocol, a shared 20% test set for all models, 10% validation within the training set for BERT early stopping, and SVM hyperparameter selection via stratified 5-fold CV on the training set, ensures a fair, apples-to-apples comparison. Second, light normalization without stopword removal preserves polarity cues (especially negations) and respects transformer pretraining, avoiding degradation often seen with aggressive text cleaning. Third, reporting per-class F1 and macro-F1 (not just accuracy) exposes performance differentials that matter in real monitoring scenarios.

There remain limitations. The neutral class is both small and noisy; even after reweighting, its F1 score plateaus around 0.26–0.27. The manual validation set (n = 139), while helpful, is modest; larger, carefully curated human labels, especially for neutral, would reduce supervision noise and provide stronger calibration anchors. Results are shown on a single, domain-specific split; although this mirrors deployment, future cross-domain or temporal tests would help assess generalization.

CONCLUSION

On the fixed 20% standard test split, IndoBERT delivers a more balanced outcome than SVM under the severe class imbalance present in Ruangguru reviews. While SVM attains the highest overall accuracy (0.888), its macro-F1 is the lowest (0.543), with F1-neutral only 0.117, indicating strong bias toward the majority (positive) class. In contrast, IndoBERT configurations consistently raise minority-class performance. The class-weighted variant (neutral boost = 1.3, LR = 3e-5, 4 epochs) yields the best macro-F1 (0.601) and F1-neutral (0.267), demonstrating that light reweighting improves balance without heavy oversampling. The baseline IndoBERT (no explicit balancing; LR = 2e-5, 3 epochs) achieves the highest IndoBERT accuracy (0.867) with macro-F1 0.596 and F1-neutral 0.255, suggesting strong pretrained representations already capture much of the sentiment signal. The sampler-only setting increases neutral recall (0.507) but reduces overall accuracy (0.852; macro-F1 0.589). These findings indicate that, for imbalanced Indonesian review streams, IndoBERT (particularly with mild class weighting) is a robust default, while SVM remains a competitive accuracy-oriented baseline. Persistent difficulty in the neutral class reflects linguistic ambiguity, label noise from rating-based supervision, and limited neutral

*name of corresponding author

examples. In practice, platforms should prioritize macro-F1 and optimized IndoBERT for monitoring minority opinions. Future work should expand manual annotations, explore focal/LDAM losses and calibrated thresholds, and evaluate domain-adaptive pretraining or targeted augmentation to further lift neutral detection without sacrificing performance on other classes.

REFERENCES

Anam, M. K., Fitri, T. A., Agustin, A., Lusiana, L., Firdaus, M. B., & Nurhuda, A. T. (2023). Sentiment Analysis for Online Learning using The Lexicon-Based Method and The Support Vector Machine Algorithm. *ILKOM Jurnal Ilmiah*, *15*(2), 290–302. https://doi.org/10.33096/ilkom.v15i2.1590.290-302

Ashbaugh, L., & Zhang, Y. (2024). A Comparative Study of Sentiment Analysis on Customer Reviews Using Machine Learning and Deep Learning. doi.org/10.3390/computers13120340

Fitri, E., Yuliani, Y., Rosyida, S., & Gata, W. (2020). Analisis Sentimen Terhadap Aplikasi Ruangguru Menggunakan Algoritma Naive Bayes , Random Forest Dan Support Vector Machine. *18*(1), 71–80. https://doi.org/10.26623/transformatika.v18i1.2317

Imtihan, K., Mutawali, L., Bagye, W., & Tantoni, A. (2025). Automated Label Extraction for Sentiment Analysis in Indonesian Text. *15*(3). https://doi.org/10.18517/ijaseit.15.3.20602

Jannah, N. Z. B., & Kusnawi, K. (2024). Comparison of Naïve Bayes and SVM in Sentiment Analysis of Product Reviews on Marketplaces. *Sinkron*, *8*(2), 727–733. https://doi.org/10.33395/sinkron.v8i2.13559

Kono, M. F., Fajri, I. N., & Pristyanto, Y. (2025). Public Sentiment Analysis on Corruption Issues in Indonesia Using IndoBERT Fine-Tuning , Logistic Regression , and Linear SVM. *9*(5), 2616–2628. https://doi.org/10.30871/jaic.v9i5.10537

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 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference*, 757–770. https://doi.org/10.18653/v1/2020.coling-main.66

Masitoh, D., Alif, A., Selamet, C., Rahmawati, S. D., Bintang, A., Setiawan, A., Studi, P., Informasi, S., & Kudus, U. M. (2025). Analisis Sentimen Ulasan Pengguna Gojek Menggunakan Metode Support Vector Machine. *6*(2), 364–372. https://doi.org/10.24127/jmsi.v6i2.9191

Munthe, S. R., & Levianti, R. A. (2025). Hybrid Framework Addressing Imbalanced Data in Indonesian E-commerce Multimodal Sentiment Analysis. *14*(November), 363–370. https://doi.org/10.34148/teknika.v14i3.1332

Mushtaha, E., Abu Dabous, S., Alsyouf, I., Ahmed, A., & Raafat Abdraboh, N. (2022). The Challenges And Opportunities Of Online Learning And Teaching At Engineering And Theoretical Colleges During The Pandemic. *Ain Shams Engineering Journal*, *13*(6), 101770. https://doi.org/10.1016/j.asej.2022.101770

Nawulansih, D. F., Santi, N. C., Aristia, I., Informatika, T., Nahdlatul, U., Sunan, U., Bojonegoro, G., Informasi, S., Nahdlatul, U., Sunan, U., & Bojonegoro, G. (2025). Analisis Sentimen Ulasan Aplikasi DANA di Google Play Store : Penerapan Support Vector Machine dan Synthetic Minority Over-sampling Technique. *5*(9), 2660–2671. https://doi.org/10.52436/1.jpti.1053

Nur, N., Aryanti, A., & Suria, O. (2025). Analisis Sentimen Terhadap Pemutusan Hubungan Kerja Di Indonesia : Komparasi INDOBERT Dengan SVM , Random Forest , Dan Decision Tree Dengan Optimasi TF - IDF *10*(2), 1158–1176. https://doi.org/10.36341/rabit.v10i2.6364

Octarini, S. P., Zakiyyah, A. Y., & Purwandari, K. (2025). Comparison of IndoBERT and SVM Algorithm to Perform Aspect Based Sentiment Analysis using Hierarchical Dirichlet Process. *7*(3), 363–370. https://doi.org/10.21512/emacsjournal.v7i3.13493

Rafiandi Andhika, F., Witanti, W., & Sabrina, P. N. (2025). Analisis Sentimen Menggunakan Metode IndoBERT pada Ulasan Aplikasi Zoom Menggunakan Fitur Ekstrasi GloVe. *9*, 2025. https://doi.org/10.47002/metik.v9i2.1098

Sarasvananda, I. B. G., Selivan, D., Radhitya, M. L., & Putra, I. N. T. A. (2022). Analisis Sentimen Pada Pembelajaran Daring Di Indonesia Melalui Twitter Menggunakan Naïve Bayes Classifier. *SINTECH (Science and Information Technology) Journal*, *5*(2), 227–233. https://doi.org/10.31598/sintechjournal.v5i2.1241

Singgalen, Y. A. (2025). Performance Analysis of IndoBERT for Sentiment Classification in Indonesian Hotel Review Data. *6*(2), 976–986. https://doi.org/10.47065/josh.v6i2.6505

Suharman, A., Sulaeman, M. K., Industri, T., Muhammadiyah, U., & Hamka, P. (2025). Analisis Sentimen Pengguna Aplikasi Livin' by Mandiri Menggunakan Metode Support Vector Machine (SVM) dengan Ekstraksi Fitur TF-IDF dan Word2Vec. *5*(8), 2201–2212. https://doi.org/10.52436/1.jpti.941

Ulinuha, A., Majid, E., & Nuari, R. (2025). Perbandingan Kinerja Metrik BERT Dan Model Machine Learning Klasik (SVM, Naive Bayes) Untuk Analisis Sentimen. *Jurnal Inovtek Polbeng*, *10*(2), 741–752. https://doi.org/10.35314/wmh3rg23

*name of corresponding author