Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16187
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Two-Stage Framework Using IndoBERT for Sentiment Analysis of Tokopedia Reviews under Extreme Class Imbalance

Ades Tikaningsih[1)*], Imam Tahyudin[2)], Berlilana[3)]
[1)2)3)]Master's Program in Computer Science, Universitas Amikom Purwokerto, Banyumas, Indonesia
[1)]adestikaningsih92@gmail.com, [2]imam.tahyudin@amikompurwokerto.ac.id, [3)]berli@amikompurwokerto.ac.id

Submitted :May 6, 2026 | **Accepted** : May 21, 2026 | **Published** : July 5, 2026

Abstract: The rapid growth of the Indonesian e-commerce industry has generated a large volume of customer reviews for sentiment analysis, but the data distribution often suffers from extreme class imbalance. The review dataset exhibits a 97.6% dominance of the positive class, causing the single-stage transformer model to produce high accuracy that does not fully represent classification capability. The baseline model achieves a macro-averaged F1-score of 0.599, with a neutral-class recall of 26.3%. Approaches based on loss function adjustment, such as class-balanced loss, focal loss, weighted cross-entropy, and decision-threshold adjustment, are unable to fundamentally address this issue, yielding only limited performance improvements. This study proposes a two-stage classification approach that decomposes the multi-class classification task into two sequential binary classification stages using a BERT-based Indonesian-language transformer model (IndoBERT). The first stage separates the positive class from the non-positive class, while the second stage distinguishes between the neutral and negative classes in a more balanced decision space. The proposed approach achieves a macro-averaged F1-score of 0.761, representing a 16.2% improvement over the baseline and outperforming all loss-function-based methods. These findings suggest that, under conditions of extreme class imbalance, simplifying the decision space through gradual task decomposition is more effective than intervention at the loss-function level. Furthermore, error propagation analysis and qualitative evaluations demonstrate that this approach improves sensitivity to minority classes, although challenges remain in cases involving ambiguous expressions.

Keywords: Class Imbalance; IndoBERT; Indonesian E-Commerce; Sentiment Analysis; Two-Stage Classification.

INTRODUCTION

E-commerce platforms in Indonesia have experienced significant growth, with transaction values reaching Rp 1,288.93 trillion in 2024, according to the 2025 Report on the Performance of Electronic Commerce (PMSE) in Indonesia published by the Central Statistics Agency (Kementerian Perdagangan, 2025). Each transaction has the potential to generate textual reviews that directly influence consumers' subsequent purchasing decisions. In a broader context, a study by YouGov (2026) found that the majority of consumers tend to abandon their purchase intentions when confronted with reviews dominated by negative sentiment, with the proportion reaching approximately 67% among consumers in the United States (Fernandes, 2026). These findings indicate that customer reviews play a crucial role in shaping consumer perceptions and decisions (Zhao et al., 2025). This impact is not limited to individual consumer behavior but also extends to seller retention, user trust, and the platform's overall revenue performance (Lumansik & Yanda Bara Kusuma, 2025). Therefore, sentiment analysis is a key component in understanding user perceptions and supporting data-driven decision-making on e-commerce platforms.

Although sentiment analysis plays a crucial role, the performance of models on Indonesian e-commerce reviews still faces a major challenge in the form of extreme and asymmetric class distributions (Munthe & Levianti, 2025). A study on the Tokopedia dataset reported a dominance of the positive class at 92.4% (Alaiya & Agusniar, 2025), while the dataset used in this study exhibits an even higher dominance, with the positive class accounting for 97.6% and the neutral and negative classes each representing only 1.2%. Under such extreme class imbalance

*Ades Tikaningsih

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

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

conditions, Transformer-based models such as IndoBERT especially when implemented without imbalance-handling techniques tend to experience a decline in minority-class performance (Setyawan & Nugraha, 2025). Approaches based on loss function modification, which have been widely used, have proven insufficient to fundamentally address this issue under conditions of extremely high imbalance ratios (Anadra et al., 2025). Furthermore, the various approaches proposed in previous studies generally focus on adjusting the loss function or data distribution but have not fully explored the effectiveness of simplifying the decision space through task decomposition strategies (Razali et al., 2025). As a result, models are required to solve complex multi-class tasks under conditions of highly imbalanced data distributions. Therefore, there remains a limited body of research that systematically applies and evaluates the effectiveness of stepwise classification strategies for addressing extreme class imbalance in sentiment analysis tasks.

To address these limitations, this study employs a task decomposition strategy through a two-stage classification approach. In the first stage, the IndoBERT model is trained as a binary classifier to distinguish between positive and non-positive reviews. This stage aims to separate the majority class from the minority classes, thereby reducing the effect of class distribution dominance during the learning process, in line with hierarchical classification and problem decomposition approaches that have been used to improve classification performance in complex scenarios (S. Lin et al., 2025). Subsequently, only reviews predicted as non-positive are forwarded to the second stage, where another IndoBERT model is trained to distinguish between neutral and negative reviews. Unlike previous approaches that focused on adjusting the loss function or data distribution (Anadra et al., 2025), this study emphasizes the effectiveness of sequential task decomposition in mitigating classification challenges under conditions of extreme class imbalance.

Based on these issues, the main contributions of this study are as follows. First, this study presents an empirical evaluation of a two-stage classification approach using a dataset of Indonesian e-commerce reviews with extreme class imbalance. Second, this study compares the effectiveness of a stepwise task decomposition strategy with that of a loss-reweighting approach in improving minority-class detection performance. Third, this study analyzes the trade-off between error propagation in the initial stage and improved discriminative power in the subsequent stage, accompanied by a qualitative analysis of classification error patterns.

LITERATURE REVIEW

Class imbalance has become one of the main challenges in text classification tasks, particularly in sentiment analysis with highly asymmetric data distributions. Studies on Transformer-based models show that high aggregate accuracy does not always reflect the model's ability to detect minority classes. For example, (Saputra & Ginting, 2025) reported that the IndoBERT model achieved an accuracy of 83%, yet its performance on minority classes remained low, with a Macro-F1 score of 0.61 and very low minority recall. This finding aligns with (Yuan et al., 2025), who demonstrated that class distribution imbalance directly reduces model performance on distribution-based metrics such as Macro-F1. A similar pattern has also been observed in RoBERTa- and BERT-based models. (He, 2024) reported that, on the imbalanced Twitter Airline dataset, models without imbalance-handling techniques achieved an accuracy of around 82%, but the F1-scores for the minority classes (positive and neutral) were only 59.31% and 65.61%, far below that of the majority class (negative), which reached 84.78%.

To address this issue, various loss-based approaches have been developed, such as class reweighting, focal loss, and cost-sensitive learning. However, the effectiveness of these methods tends to be limited to conditions of extreme imbalance. (Cartus et al., 2023) report that the effectiveness of loss-based approaches decreases significantly when the imbalance ratio exceeds 100:1. On the other hand, the classification structure also affects a model's ability to handle class imbalance. In single-stage approaches, a model must distinguish all classes simultaneously within a single decision space, which can intensify competition among classes and reinforce bias toward the majority class (Kyritsis et al., 2025). Studies by (Leevy et al., 2018) and (Henning et al., 2023) show that multi-class classification under imbalanced conditions is inherently more difficult than binary classification approaches, as models fail to establish adequate decision boundaries and tend to favor the majority class within the same decision space.

Alternatively, task-decomposition-based approaches, such as cascaded classification and two-stage classification, are increasingly being used to simplify the decision space by breaking down the classification task into several simpler stages. This approach allows for the separation of majority class detection and minority class discrimination, enabling the model to learn more stable decision boundaries on a more balanced subset of data (W. Zhou et al., 2024). However, the application of decomposition approaches to sentiment analysis with extreme class imbalance remains limited and has not yet been systematically compared with loss-function-based approaches. A study by (S. Cui et al., 2023) applied a two-stage ensemble approach to social media sentiment classification and achieved competitive results; however, that study was conducted on a dataset with a relatively balanced class distribution and did not include comparisons with loss-based approaches such as focal loss or class-weighted loss.

*Ades Tikaningsih
This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16187
e-ISSN : 2541-2019
p-ISSN : 2541-044X

METHOD

This study employs an experimental approach to compare the performance of several strategies for addressing extreme class imbalance in sentiment classification tasks. The methodology includes data preparation, the development of a transformer-based baseline model, the implementation of loss-function-based approaches, and the evaluation of two-stage classification approaches. The evaluation focuses on the Macro-F1 metric to ensure balanced performance across all classes.

Data Preprocessing

This study uses data from the Kaggle platform, specifically the Tokopedia Product Reviews 2025 dataset (Abdurrahman, 2025). This dataset consists of 65,543 Tokopedia product reviews collected via web scraping, focusing on products with high transaction activity across various attributes. The dataset uses the review_text attribute as the text input and sentiment_label as the classification target, which is then encoded into three classes: positive (0), neutral (1), and negative (2). The class distribution in the dataset shows extreme imbalance. The positive class dominates at 97.6%, while the neutral and negative classes account for only 1.2% and 1.2%, respectively. This asymmetry has the potential to introduce model bias toward the majority class and produce unrepresentative evaluations when relying solely on accuracy metrics (Widyananda & Fauzi, 2025).

The data is processed in its original text form without applying conventional preprocessing techniques such as stemming or stopword removal to preserve contextual and semantic information, as the IndoBERT model can learn feature representations autonomously through self-attention mechanisms without relying on feature engineering (Adhim & Cahyono, 2025). Next, the text is tokenized using the indobenchmark/indobert-base-p1 tokenizer, with padding and truncation applied to a maximum length of 128 tokens. This value was determined based on the token length distribution in the dataset, where 99.78% of reviews are under 128 tokens long, with an average of 18 tokens and a median of 14 tokens. The short nature of these reviews allows for the use of a 128-token limit to optimally represent information while maintaining computational efficiency (Z. G. Zhou, 2022).

As the final step in data preparation, the dataset is divided into training (80%), validation (10%), and testing (10%) sets. This division is performed using stratified sampling to preserve class proportions across all subsets.

Baseline Transformer Model

The baseline model uses the IndoBERT architecture (indobenchmark/indobert-base-p1), which has been fine-tuned for a three-class sentiment classification task (positive, neutral, negative). Given a review x, the model generates a contextual representation using the special token $[CLS]$, which is then projected into the output class space using a linear layer and a softmax function (Wilie et al., 2020) :

$$P(y|x) = softmax(Wh_{CLS} + b) \quad (1)$$

Here, $y \in \{0,1,2\}$ corresponds to positive, neutral, and negative sentiment, respectively. Under conditions of highly imbalanced data distribution, this approach tends to produce bias toward the majority class, resulting in poor performance on minority classes.

Loss-Based Approaches for Class Imbalance

Class-Balanced Loss

This work adopts the Class-Balanced Loss framework proposed by (Y. Cui et al., 2019) as a solution for extreme class imbalance, using class weights computed from the effective number of samples principle.

$$\mathcal{L}_{CB} = -\sum_{i=1}^{C} w_i y_i \log(\hat{y}_i) \quad (2)$$

In this study, the β value was set to 0.9999 in accordance with the recommendations of (Y. Cui et al., 2019) and adjusted to the characteristics of the dataset, which exhibits a highly imbalanced class distribution with significant dominance of the majority class.

Focal Loss

According to (T. Lin et al., 2017), Focal Loss was developed as a variation of the cross-entropy loss function aimed at addressing class imbalance by down-weighting easily classified instances. Mathematically, the Focal Loss function is formulated as follows:

$$FL(pt) = -a_t(1 - p_t)^{\gamma} \log(p_t) \quad (3)$$

where, p_t is the model's predicted probability for the target class. The parameter a_t serves as a weighting factor between classes, while the parameter $\gamma \geq 0$ (focusing parameter) controls the degree to which the loss contribution of easily classified samples is reduced.

*Ades Tikaningsih

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16187
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Weighted Cross-Entropy

Weighted Cross-Entropy (WCE) is a modification of the cross-entropy loss function that assigns different weights to each class to address class distribution imbalance. The WCE loss function is defined as follows:

$$WCE(pt) = -w_t \log(p_t) \quad (4)$$

where w_t is the class weight, which is typically calculated as the inverse of class frequency in the dataset. This approach aims to increase the model's sensitivity to minority classes by increasing the penalty for prediction errors in those classes (Siva & Hoki, 2026).

Threshold Adjustment Strategy

The threshold tuning strategy is applied as a post-processing approach to improve the model's sensitivity to minority classes by adjusting decision thresholds based on the validation set. In the classification setting, label predictions are determined as follows:

$$\hat{y} = \begin{cases} y_m, & if\ P(y_m|x) \geq \tau \\ \arg\max_{y \neq y_m} P(y|x), & \text{otherwise} \end{cases} \quad (5)$$

where y_m represents the neutral and negative classes. The value of τ is determined through a search procedure on the validation set with the aim of improving metrics such as F1-score or recall for minority classes.

Proposed Hierarchical Two-Stage Transformer

This study proposes a hierarchical two-stage transformer approach as a decomposition-based strategy for addressing sentiment classification under conditions of extreme class imbalance by decomposing the three-class sentiment classification problem into two simpler binary classification subproblems. In general, multi-class classification approaches model the class distribution as follows (Oh, 2022):

$$P(y \mid x), y \in \{0,1,2\} \quad (6)$$

This formulation requires the model to learn decision boundaries simultaneously across all classes (Moral et al., 2022). Under conditions of highly imbalanced class distributions, this process increases class competition, resulting in decision boundaries biased toward the majority class (Abdulkreem & Abdulazeez, 2025). The proposed approach models the classification process as a sequential decomposition into two simpler binary classification decisions.

$$P(y \mid x) = P(c_1|x)\, P(c_2|x, c_1) \quad (7)$$

where $c_1$represents the first stage decision (positive or non-positive), and c_2 represents the second-stage classification (neutral or negative), which is applied only when c_1 is classified as non-positive. This approach is related to the concepts of classifier decomposition and hierarchical classification, in which complex predictions are decomposed into a series of simpler conditional decisions (Murphy, 2022).

To simplify the decision space under extreme class imbalance, the proposed framework decomposes the classification process into two sequential stages, as illustrated in Fig. 1.

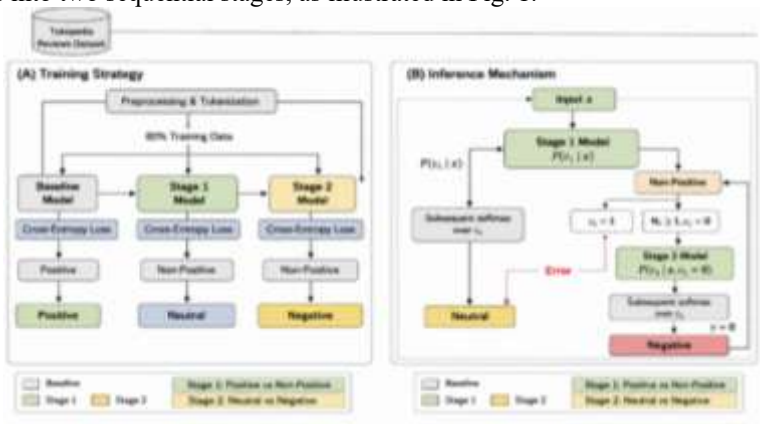

Fig. 1 Overview of the proposed two-stage framework. The left side shows independent training, while the right side illustrates sequential inference.

Stage-1: Positive vs Non-Positive

In the first stage, the original labels are transformed into binary variables as follows:

$$c_1 = \begin{cases} 0, if\ y = 0 (positive) \\ 1, if\ y \in \{1,2\} (non-positive) \end{cases} \quad (8)$$

The first-stage model performs the following classification task:

*Ades Tikaningsih

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16187
e-ISSN : 2541-2019
p-ISSN : 2541-044X

$$P(c_1 \mid x) \tag{9}$$

The model is optimized using the binary cross-entropy loss function:

$$\mathcal{L}_1 = -y \log(\hat{y}) + (1-y)\log(1-\hat{y}) \tag{10}$$

The choice of a positive-versus-non-positive separation scheme is not arbitrary. In the context of e-commerce review data, the positive class generally dominates the data distribution; therefore, separating it at an early stage allows the model to first identify the dominant class before performing more refined discrimination among minority classes. This strategy aligns with the coarse-to-fine classification principle in the literature on hierarchical classification and cascaded models (Xiong & Yao, 2023).

Stage-2: Neutral vs Negative

The second stage is trained using a subset of the training data with $c_1 = 1$, samples belonging to the non-positive class. The labels for this stage are defined as follows:

$$c_2 = \begin{cases} 0, if\ y = 1\ (neutral) \\ 1, if\ y = 2\ (negative) \end{cases} \tag{11}$$

The second-stage model learns the conditional distribution:

$$P(c_2 \mid x, c_1 = 1) \tag{12}$$

The model is optimized using the binary cross-entropy loss function:

$$\mathcal{L}_2 = -y \log(\hat{y}) + (1-y)\log(1-\hat{y}) \tag{13}$$

The data subset for the second stage was created based on the ground truth labels in the training data to prevent inter-stage data leakage. Consequently, the second-stage training process is not affected by prediction errors from the first stage.

End-to-End Inference Mechanism

During the inference stage, the final prediction is obtained through a sequential classification mechanism that combines two binary models sequentially. This mechanism reflects the decomposition of the classification decision into two simpler stages, where the second-stage decision is conditional on the first-stage result. Given a review x, the first-stage model generates a prediction:

$$\widehat{c_1} = arg\ max\ P(c_1 \mid x) \tag{14}$$

If $\widehat{c_1} = 0$ (positive), then the final prediction is immediately set to:

$$\hat{y} = 0 \tag{15}$$

Conversely, if $\widehat{c_1} = 1$(non-positive), the sample is passed to the second stage to distinguish between the neutral and negative classes. The second-stage prediction is defined as follows:

$$\widehat{c_2} = arg\ max\ P(c_2 \mid x, c_1 = 1) \tag{16}$$

Therefore, the final prediction is determined as:

$$\hat{y} = \begin{cases} 0, & if\ \widehat{c_1} = 0 \\ 1, & if\ \widehat{c_1} = 1 \wedge \widehat{c_2} = 0 \\ 2, & if\ \widehat{c_1} = 1 \wedge \widehat{c_2} = 1 \end{cases} \tag{17}$$

However, this approach inherently introduces the potential for error propagation across stages (Wu et al., 2024) because errors in the first stage are critical, as samples predicted as positive are not passed to the second stage, meaning such errors cannot be corrected by the second-stage model. In this context, the system's overall performance is determined not only by the accuracy of each model independently but also by interactions between stages within the inference pipeline. This means that end-to-end performance cannot be assumed to reflect a simple combination of the independently evaluated performance of each stage. Furthermore, the data distribution received by the second stage during inference depends on the output of the first stage, resulting in a distribution that differs from the training-time data distribution, which uses a subset based on ground truth labels (Fu et al., 2021).

Experimental Setup

All experiments used the IndoBERT architecture (indobenchmark/indobert-base-p1) with a consistent fine-tuning procedure to ensure fair comparisons between methods. The experiments differed only in the loss function and the formulation of the classification task. All models were trained with a maximum sequence length of 128, a batch size of 16, a learning rate of 2e-5, and the AdamW optimizer with a weight decay of 0.01. The fine-tuning process was conducted over three epochs using a consistent configuration across all experiments, including the baseline, loss-based approaches, and the two-stage framework.

*Ades Tikaningsih

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16187
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Evaluation and Analysis Metrics

Model performance was evaluated using multi-class classification metrics that account for class distribution imbalance. All metrics were calculated based on the confusion matrix for each class $c \in \{0,1,2\}$, where the labels represent positive (0), neutral (1), and negative (2) sentiments.

Accuracy is defined as the percentage of correct predictions out of all samples within the dataset (Sa, 2025):

$$Accuracy = \frac{\sum_{c=1}^{C} TP_c}{N} \tag{18}$$

Precision refers to the proportion of accurate predictions among all outputs assigned to class c. On the other hand, recall assesses the fraction of class c samples that the model correctly classifies (Rizqilla et al., 2025)

$$Precision_c = \frac{TP_c}{TP_c + FP_c} \tag{19}$$

$$Recall_c = \frac{TP_c}{TP_c + FN_c} \tag{20}$$

As the primary metric, this study uses the Macro F1-score, which represents the unweighted average of F1-scores across all classes. The Macro F1-score was chosen because it assigns equal weight to each class, making it more appropriate for evaluating model performance under class imbalance conditions (Ribeiro, 2015).

$$F1_c = 2 \times \frac{Precision_c \times Recall_c}{Precision_c + Recall_c} \tag{21}$$

$$Macro\ F1 = \frac{1}{C}\sum_{C=1}^{C} F1_c \tag{22}$$

RESULT

Dataset Characteristics and Baseline Performance

Table 1 presents the performance of the baseline IndoBERT model using the standard cross-entropy loss function on the three-class sentiment classification task.

Table 1. Performance Metrics of the Baseline Model

Class	Precision	Recall	F1
Positive	0.988	0.994	0.991
Neutral	0.350	0.263	0.300
Negative	0.581	0.450	0.507
Accuracy		**0.978**	
Macro F1		**0.599**	

The baseline model achieved an accuracy of 97.8% but attained only a Macro F1-score of 59.9%, reflecting an imbalance in performance across classes. Further analysis shows that performance on the minority classes remains low, with recall for the neutral and negative classes at 26.3% and 45.0%, respectively. In contrast, recall for the positive class reaches 99.4%, suggesting a strong bias toward the majority class. These findings confirm that the use of the standard cross-entropy loss function is insufficient to handle extreme class imbalance, necessitating alternative approaches (Zahro et al., 2025).

Results of Loss-Based and Threshold-Based Interventions

To address class imbalance, we evaluated several loss-based and threshold-based approaches. Table 2 presents a comparison of their performance.

Table 2. Performance Comparison of Loss-Based and Threshold-Based Interventions

Model	Accuracy	Macro F1	Recall (Neutral)	Recall (Negative)
Baseline	0.978	0.599	0.263	0.450
Class-Balanced Loss	0.976	0.591	0.200	0.550
Focal Loss	0.977	0.595	0.262	0.450
Weighted CE	0.977	0.593	0.275	0.412
Threshold Moving	0.980	0.595	0.290	0.410

*Ades Tikaningsih

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16187
e-ISSN : 2541-2019
p-ISSN : 2541-044X

The results show that all intervention methods failed to outperform the baseline on aggregate metrics, with the Macro F1-score remaining stagnant and even declining slightly from 0.599 to the range of 0.591–0.595. Per-class recall analysis revealed a consistent trade-off: methods that improved recall for one minority class tended to degrade performance for other minority classes. Weighted Cross-Entropy and Threshold Moving showed improvements in neutral-class recall (0.275 and 0.290, representing improvements of 4.5% and 10.3% over the baseline, respectively), but experienced significant declines in negative-class recall (0.412 and 0.410, representing decreases of 8.4% and 8.9% from the baseline). Conversely, Class-Balanced Loss improved negative-class recall to 0.550 (+22.2%) but reduced neutral-class detection performance to 0.200 (-24.0% from the baseline).

This limitation suggests that, under conditions of extreme imbalance, interventions at the loss-function level are unable to fundamentally alter the model's bias toward the majority class. The dominance of the gradient from the positive class creates optimization pressure that is sufficiently strong that reweighting efforts through the loss function merely shift bias from one minority class to another without improving the model's overall discriminative power.

Two-Stage Framework Analysis

The evaluation of the two-stage approach was conducted in stages to understand the contribution of each component to overall performance. The analysis focused on three main aspects: (1) the ability of Stage 1 to separate the majority class (positive) from the non-positive class, (2) the ability of Stage 2 to distinguish between the neutral and negative classes in the filtered subset, and (3) the impact of integrating both stages on end-to-end performance.

Stage 1: Positive vs. Non-Positive Classification

Table 3 presents the performance of the Stage 1 model in distinguishing the majority class (positive) from the combined minority classes (non-positive).

Table 3. Stage-1 Performance Breakdown

Class	Precision	Recall	F1
Positive	0.990	0.990	0.990
Non-Positive	0.601	0.594	0.597
Accuracy		**0.980**	
Macro F1		**0.794**	

The Stage 1 model achieved a Macro F1-score of 0.794 with a recall of 0.594 for the non-positive class. Although this recall indicates that approximately 40.6% of minority samples were not successfully passed to the next stage, this performance is still better than that of the direct multi-class approach in the baseline, where recall for minority classes ranged from 0.263 to 0.450. This stage functions as a coarse filtering mechanism that isolates the dominant class before more refined discrimination is performed.

Stage 2: Neutral vs. Negative Classification

The performance of Stage 2, presented in Table 4, was evaluated using non-positive samples selected based on ground truth labels. Therefore, these results represent an ideal scenario in which all non-positive samples are assumed to have been successfully classified by Stage 1.

Table 4. Stage-2 Performance Breakdown

Class	Precision	Recall	F1
Neutral	0.693	0.763	0.726
Negative	0.736	0.663	0.697
Accuracy		**0.712**	
Macro F1		**0.712**	

The Stage 2 model achieved a Macro F1-score of 0.712 on the non-positive subset, significantly higher than the baseline performance for both minority classes in the multi-class setting (F1-scores of 0.300 and 0.507, respectively). This improvement indicates that, by eliminating the dominance of the positive class, the model can learn a more discriminative decision boundary between neutral and negative classes.

The sequential integration of the two stages yields an end-to-end Macro F1-score of 0.761, representing a significant improvement over both the baseline (0.599) and loss-function-based approaches. These results indicate that, despite sample loss in Stage 1, the combination of the two stages still improves the balance of performance across classes.

*Ades Tikaningsih

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16187
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Overall Performance Comparison

Table 5 presents a comparison of all methods evaluated in this study, including the baseline model, loss-based interventions, threshold-adjustment approaches, and the proposed two-stage framework.

Table 5. Overall Performance Comparison Across All Methods

Model	Accuracy	Macro F1
Baseline (IndoBERT)	0.978	0.599
Class-Balanced Loss	0.976	0.591
Focal Loss	0.977	0.595
Weighted CE	0.977	0.593
Threshold Moving	0.980	0.595
Two-Stage	0.980	0.761
Improvement		+16.2%

The results show that the two-stage approach achieved the highest Macro F1-score of 0.761, outperforming all loss-reweighting-based methods by a significant margin (16.2% higher than the baseline and 16.6% higher than the best loss-based method). This improvement far exceeds the performance variation produced by loss-based approaches, which varied only within a range of ±0.01. These findings indicate that the single-stage approach has fundamental limitations in handling highly imbalanced class distributions, as all classes are learned simultaneously within a single decision space. In contrast, the two-stage approach explicitly decomposes the classification task into two simpler subproblems, thereby reducing direct competition between the majority and minority classes.

To validate that the performance improvement stems from the decomposition strategy used, an ablation study was conducted by reversing the order of the classification decomposition. In the alternative strategy, Stage 1 separates the negative class from the non-negative class, while Stage 2 distinguishes between the positive and neutral classes. The results in Table 6 show that this strategy achieved a Macro-F1 score of only 0.553, which is significantly lower than that of the proposed approach. This performance drop indicates that the effectiveness of hierarchical classification heavily depends on how the decomposition is performed. Separating the majority class at an early stage proves more effective in simplifying the decision space and facilitating more stable discrimination between minority classes.

Table 6. Ablation Study on Decomposition Strategy

Strategy	Stage-1	Stage-2	Macro-F1
Proposed	Positive vs Non-Positive	Neutral vs Negative	0.761
Ablation	Negative vs Non-Negative	Positive vs Neutral	0.553

Statistical Significance Testing

The experiment was repeated using three different random seeds to evaluate the stability of the model's performance. Table 7 shows that the two-stage approach consistently yielded a higher Macro-F1 score than the baseline across all test seeds. The baseline achieved an average Macro-F1 score of 0.600 ± 0.009, whereas the two-stage approach achieved an average score of 0.771 ± 0.014. To test the statistical significance of this improvement, a paired t-test was performed, yielding a p-value of 0.0085, $p < 0.05$, and a t-statistic of 10.748. These results indicate that the improvement in the two-stage model's performance is statistically significant and is unlikely to be due to random seed variation.

Table 7. Statistical Significance Across Random Seeds

Model	Seed 42	Seed 123	Seed 777	Mean ± Std
Baseline	0.588	0.607	0.605	0.600 ± 0.009
Two-Stage	0.790	0.757	0.767	0.771 ± 0.014

Error Propagation and Interpretability

The two-stage approach inherently introduces an error propagation mechanism, whereby errors in Stage 1 directly affect the distribution of data passed to Stage 2. Based on the evaluation results, Stage 1 was able to retain approximately 59.4% of the non-positive samples, while approximately 40.6% were not passed to the next stage. Nevertheless, samples successfully processed by Stage 2 are classified with greater discriminative accuracy,

*Ades Tikaningsih

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16187
e-ISSN : 2541-2019
p-ISSN : 2541-044X

indicating a trade-off between data retention and classification quality. The difference in error patterns between the baseline model and the two-stage model can be observed through the confusion matrix in Fig. 2.

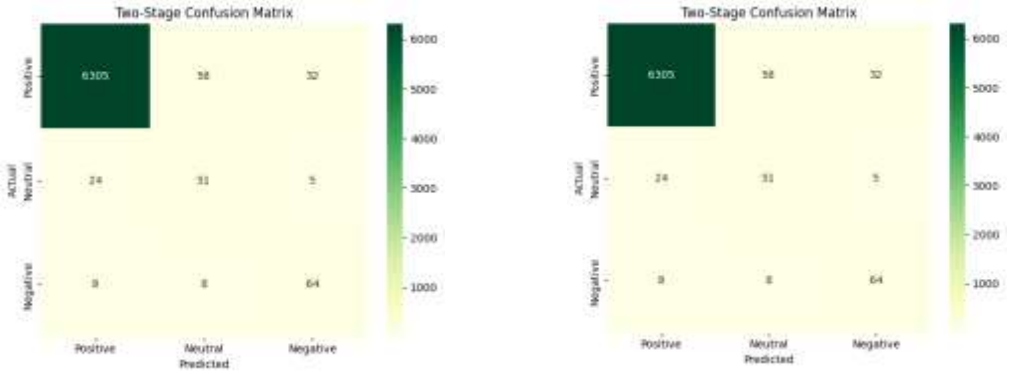

Fig. 2 Confusion Matrix Comparison Between Baseline and Two-Stage Models

The proposed two-stage approach successfully mitigated this bias, increasing recall for the neutral class from 0% to 66.7%. These results demonstrate that, by separating the decision space at an early stage, the model can capture more subtle linguistic patterns associated with previously overlooked minority-class samples. Table 6 presents several representative examples illustrating the differences in predictions between the baseline model and the two-stage approach.

Table 8. Qualitative Error Analysis

Model	True	Baseline	Two-Stage	Category Status
“sayang banget telurnya busuk”	Negative	Neutral	Negative	explicit complaint
“packaging agak kurang”	Neutral	Positive	Neutral	implicit sentiment
“barang datang kondisi rusak”	Neutral	Negative	Neutral	over-prediction
“lumayan harga dan produk rapih”	Neutral	Positive	Positive	Ambiguous
“barang cepat sampai, bagus”	Neutral	Positive	Positive	Ambiguous

The analysis in Table 8 shows that the two-stage approach consistently improves classification performance for texts containing implicit negative or neutral sentiment. The baseline model tends to classify such texts as positive, reflecting bias toward the majority class. In contrast, the two-stage approach is better at capturing weak linguistic signals, such as mild complaints or indications of product dissatisfaction. These findings confirm that the performance limits of the two-stage framework are largely determined by the intrinsic semantic ambiguity of review texts rather than limitations in the model architecture.

To gain a deeper understanding of the model’s behavior, we conducted an interpretability analysis using LIME, as shown in Fig. 3.

Fig. 3 LIME-Based Token Importance Analysis for Minority-Class Prediction

The interpretability results show that the model utilizes complaint-related tokens such as “lama,” “kurang,” and “packing,” as well as contrastive expressions such as “tapi,” during the sentiment prediction process. These findings indicate that the two-stage approach is capable of capturing minority-class linguistic patterns and evaluative contexts that are often overlooked by single-stage approaches under highly imbalanced class distributions.

DISCUSSIONS

The application of decomposition via two-stage classification consistently outperforms loss-function-adjustment-based methods in handling extreme class imbalance. At extremely high imbalance ratios (>80:1), the gradient contribution from the majority class becomes overly dominant, causing reweighting efforts at the loss-function level to merely shift bias among minority classes rather than improve overall discriminative ability. The

*Ades Tikaningsih

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

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

two-stage architecture circumvents this fundamental limitation, yielding a Macro F1-score of 0.761 (16.2% improvement over baseline) demonstrating that task decomposition, not loss-function modification, is the critical mechanism for handling extreme imbalance.

However, the two-stage approach introduces an unavoidable trade-off: error propagation between stages. Analysis shows that Stage 1 successfully passes only about 59.4% of non-positive samples to the next stage, with the remaining 40.6% incorrectly classified as positive and therefore unrecoverable in subsequent stages. Nevertheless, samples successfully passed by Stage 1 were classified by Stage 2 with a Macro F1-score of 0.712, far surpassing the baseline performance on the same subset. These findings indicate that the benefit of simplifying the decision space in Stage 2 is empirically greater than the cost of losing samples in Stage 1, resulting in superior end-to-end performance overall. Qualitative analysis further supports these findings: the two-stage approach successfully corrected four out of five baseline failure cases, particularly for texts containing explicit complaints and implicit sentiment. However, two cases involving ambiguous expressions still failed in both models, indicating that the remaining performance limitations are influenced more by the intrinsic semantic ambiguity of the Indonesian language than by limitations of the classification strategy.

There are several limitations that must be acknowledged in this study. First, all experiments were conducted on a single dataset from a single platform (Tokopedia), so the generalizability of the findings to other e-commerce contexts or domains requires further validation. Second, the two-stage approach is highly dependent on the quality of decisions made in the initial stage, meaning that errors in Stage 1 are irreversible within the inference pipeline. Third, intrinsic linguistic challenges such as ambiguous expressions and mixed sentiments within a single sentence cannot yet be addressed through modifications to the classification strategy alone and may require additional approaches such as generative-model-based data augmentation or label refinement through inter-annotator agreement.

CONCLUSION

Extreme class imbalance in multi-class sentiment classification is a major challenge that causes models to become biased toward the majority class and fail to detect minority classes effectively. Under these conditions, metrics such as accuracy become unrepresentative, necessitating approaches capable of handling imbalanced data distributions while maintaining the ability to distinguish between classes. This study demonstrates that, under conditions of extremely dominant majority classes, all loss-function-based intervention approaches tested including Class-Balanced Loss, Focal Loss, Weighted Cross-Entropy, and Threshold Moving proved unable to fundamentally resolve this issue, as changes in Macro F1-score remained within ±0.01 of the baseline due to the overwhelming dominance of gradients from the majority class. This study implements and evaluates a two-stage classification strategy that decomposes the multi-class task into two sequential binary decisions. Stage 1 separates the positive class from the non-positive class as an initial filtering mechanism, while Stage 2 performs more precise discrimination between neutral and negative classes in a more balanced decision space. This approach yields an end-to-end Macro F1-score of 0.761, representing a 16.2% improvement over the baseline and significantly outperforming all loss-based methods, despite a trade-off involving unrecoverable error propagation from Stage 1 amounting to 40.6%. These findings indicate that, under conditions of extreme class imbalance, simplifying the decision space through sequential task decomposition provides a more effective solution than interventions at the loss-function level alone.

REFERENCES

Abdulkreem, R. Z., & Abdulazeez, A. M. (2025). A Comparative Study of Multi-class Classification Based on Imbalanced Data: A Review. The Indonesian Journal of Computer Science, 14(5), 8211–8233. https://doi.org/https://doi.org/10.33022/ijcs.v14i5.5020

Abdurrahman, S. (2025). Tokopedia Product Reviews 2025. Kaggle.Com. https://doi.org/10.34740/KAGGLE/DS/8992580

Adhim, N. F., & Cahyono, N. (2025). Optimization of IndoBERT for Sentiment Analysis of FOMO on Social Media Through Fine-Tuning and Hybrid Labeling. Journal of Applied Informatics and Computing (JAIC), 9(6), 3786–3797. https://doi.org/https://doi.org/10.30871/jaic.v9i6.11686

Alaiya, A., & Agusniar, C. (2025). Sentiment Analysis of E-Commerce Product Reviews on Tokopedia Using Support Vector Machine. Journal of Applied Informatics and Computing, 9(5), 2869–2878. https://doi.org/https://doi.org/10.30871/jaic.v9i5.10977

Anadra, R., Wijayanto, H., & Sadik, K. (2025). Sentiment Analysis of Tokopedia Customer Reviews Using BiLSTM and IndoBERT with Comparative Analysis of Preprocessing and Labeling Methods. International Journal of Advances in Data and Information Systems, 6(3), 773–788. https://doi.org/10.59395/ijadis.v6i3.1458

*Ades Tikaningsih

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

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

Cartus, A. R., Samuels, E. A., Cerdá, M., & Marshall, B. D. L. (2023). Outcome class imbalance and rare events: An underappreciated complication for overdose risk prediction modeling. National Library of Medicine, 118(6), 1167–1176. https://doi.org/https://doi.org/10.1111/add.16133

Cui, S., Han, Y., Duan, Y., Li, Y., Zhu, S., & Song, C. (2023). A Two-Stage Voting-Boosting Technique for Ensemble Learning in Social Network Sentiment Classification. MDPI, 25, 1–24. https://doi.org/10.3390/e25040555

Cui, Y., Jia, M., Lin, T., & Tech, C. (2019). Class-Balanced Loss Based on Effective Number of Samples. https://doi.org/https://doi.org/10.48550/arXiv.1901.05555

Fernandes, J. (2026). How do American consumers verify online retailers before purchasing? Yougov.Com. https://yougov.com/en-us/articles/53837-how-do-american-consumers-verify-online-retailers-before-purchasing

Fu, Y., Huang, H., Guan, Y., Wang, Y., & Liu, W. (2021). Knowledge-Based Systems EBRB cascade classifier for imbalanced data via rule weight updating. Knowledge-Based Systems, 223, 107010. https://doi.org/10.1016/j.knosys.2021.107010

He, L. (2024). Enhanced twitter sentiment analysis with dual joint classifier integrating RoBERTa and BERT architectures. Frontiers in Physics, December, 1–12. https://doi.org/10.3389/fphy.2024.1477714

Henning, S., Beluch, W., Fraser, A., & Friedrich, A. (2023). A Survey of Methods for Addressing Class Imbalance in Deep-Learning Based Natural Language Processing. Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, 2022, 523–540.

Kementerian Perdagangan. (2025). Kinerja Perdagangan Melalui Sistem Elektronik (PMSE) Indonesia Tahun 2025.

Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A., & Seliya, N. (2018). A survey on addressing high - class imbalance in big data. Journal of Big Data. https://doi.org/10.1186/s40537-018-0151-6

Lin, S., Frasincar, F., & Klinkhamer, J. (2025). Hierarchical Deep Learning for Multi-label Imbalanced Text Classification of Economic Literature. Applied Soft Computing, 31(0).

Lin, T., Ai, F., & Doll, P. (2017). Focal Loss for Dense Object Detection. https://doi.org/10.48550/arXiv.1708.02002

Lumansik, D. K., & Yanda Bara Kusuma. (2025). THE INFLUENCE OF ONLINE CUSTOMER REVIEW AND ONLINE CUSTOMER RATINGS ON PURCHASE DECISIONS IN TIKTOK SHOP. Indonesian Interdisciplinary Journal of Sharia Economics (IIJSE), 8(2), 4427–4443. https://doi.org/https://doi.org/10.31538/iijse.v8i2.6108

Moral, P. D. E. L., Nowaczyk, S., & Pashami, S. (2022). Why Is Multiclass Classification Hard ? IEEE Access, 10(June). https://doi.org/10.1109/ACCESS.2022.3192514

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

Murphy, K. P. (2022). Adaptive Computation and Machine Learning.

Oh, S. H. (2022). Probabilistic Target Encoding of Neural Networks with Softmax Output Nodes. International Journal of Contents, 18(4). https://doi.org/https://doi.org/10.5392/IJoC.2022.18.4.102

Razali, M. N., Arbaiy, N., & Lin, P. (2025). Optimizing Multiclass Classification Using Convolutional Neural Networks with Class Weights and Early Stopping for Imbalanced Datasets. MDPI, 14, 1–14. https://doi.org/https://doi.org/10.3390/electronics14040705

Ribeiro, R. P. (2015). A Survey of Predictive Modelling under Imbalanced Distributions. https://doi.org/https://doi.org/10.48550/arXiv.1505.01658

Rizqilla, M., Nurani, D., Supriatin, Rahmi, A. N., & Maemunah, M. (2025). Analisis Sentimen Kinerja Jokowi Menggunakan Metode Naïve Bayes. Riset Dan E-Jurnal Manajemen Informatika Komputer, 9(2), 550–561. https://doi.org/http://doi.org/10.33395/remik.v9i2.14670

Sa, M. Ü. (2025). Accuracy Comparison of Machine Learning Algorithms on World Happiness Index Data. MDPI, 13. https://doi.org/https://doi.org/10.3390/math13071176

Saputra, I., & Ginting, G. L. (2025). Sentiment Analysis of Tokopedia Customer Reviews using IndoBERT and SMOTE for Class Imbalance Handling. Journal of Computer System and Informatics (JoSYC), 7(1), 20–27. https://doi.org/10.47065/josyc.v7i1.8748

Setyawan, A. F., & Nugraha, R. I. (2025). OPTIMIZING GPT AND INDOBERT FOR SENTIMENT ANALYSIS AND CONSUMER TREND PREDICTION ON LAZADA PRODUCT REVIEWS. JIKO (Jurnal Informatika Dan Komputer), 8(2), 113–120. https://doi.org/10.33387/jiko.v8i2.10066

*Ades Tikaningsih

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

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

Function in Convolutional Neural Network for Acne Image Classification. Journal of Artificial Intelligence and Engineering Applications, 5(2), 2–7.

Siva, A. B., & Hoki, L. (2026). Comparison of IndoBERT and SVM Performance in Sentiment Analysis of Digital Education Platforms. Sinkron : Jurnal Dan Penelitian Teknik Informatika, 10(1), 64–74. https://doi.org/10.33395/sinkron.v10i1.15472

Widyananda, W., & Fauzi, A. (2025). Sentiment analysis of Indonesian e-commerce product reviews using machine learning classifiers. International Journal of Science and Research Archive, 16(02), 445–450. https://doi.org/https://doi.org/10.30574/ijsra.2025.16.2.2345

Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., & Purwarianti, A. (2020). IndoNLU : Benchmark and Resources for Evaluating Indonesian Natural Language Understanding. Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processin, 843–857.

Wu, Z., Wang, Y., Shen, L., Hu, F., & Yu, H. (2024). Leveraging Uncertainty for Depth-Aware Hierarchical Text Classification. Computers, Materials and Continua. https://doi.org/10.32604/cmc.2024.054581

Xiong, H., & Yao, A. (2023). Deep Imbalanced Regression via Hierarchical Classification Adjustment. https://doi.org/https://doi.org/10.48550/arXiv.2310.17154

Yuan, D., Liang, G., Liu, B., & Liu, S. (2025). Qwen TextCNN and BERT models for enhanced multilabel news classification in mobile apps. Nature, 15, 1–22. https://doi.org/https://doi.org/10.1038/s41598-025-27497-6

Zahro, A. C., Alzami, F., Sani, R. R., & Fahmi, A. (2025). Fairer Public Complaint Classification on LaporGub : Integrating XLM-RoBERTa with Focal Loss for Imbalance Data. Sinkron : Jurnal Dan Penelitian Teknik Informatika, 9(4), 1850–1862. https://doi.org/10.33395/sinkron.v9i4.15260

Zhao, M., Wang, S., & Id, T. X. (2025). The social welfare effect of e-commerce product reputation information asymmetry from the perspective of network externality. PLoS ONE, 1–20. https://doi.org/10.1371/journal.pone.0313852

Zhou, W., Liu, C., Yuan, P., & Jiang, L. (2024). applied sciences An Undersampling Method Approaching the Ideal Classification Boundary for Imbalance Problems. Applied Sciences, 14. https://doi.org/https://doi.org/10.3390/app14135421

Zhou, Z. G. (2022). Research on Sentiment Analysis Model of Short Text Based on Deep Learning. Scientific Programming, 2022. https://doi.org/10.1155/2022/2681533

*Ades Tikaningsih

This is anCreative Commons License This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.