Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16209
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Comparative Evaluation of IndoBERT-Based Architectures for Imbalanced Indonesian News Title Classification

Erwin Sirait[1)*], Juni Ismail[2)]
[1)]Department of Computerized Accounting, Politeknik Bisnis Indonesia, Simalungun, Indonesia
[2)]Department of Computer Engineering, Politeknik Bisnis Indonesia, Simalungun, Indonesia
[1)*]esirait946@gmail.com, [2)]juniismaillll@gmail.com

Submitted: May 13, 2026 | **Accepted:** Jun 29, 2026 | **Published:** July 14, 2026

Abstract: Stacking multiple imbalance-mitigation techniques on top of a pretrained transformer is widely assumed to compound their individual benefits, yet rigorous component-wise evidence for this assumption remains scarce in the Indonesian text classification literature. Four classification architectures are compared in this work on a publicly available Indonesian news title corpus. The working set contains 27,266 short headlines, drawn as a 30% stratified subsample from a cleaned corpus of 90,891 headlines, spread over nine target categories with a class ratio of 13.29. Three reference architectures are constructed: an LSTM trained from scratch with Random Oversampling, a bidirectional LSTM augmented with additive attention, and a fine-tuned IndoBERT on the oversampled training partition. A fourth architecture extends IndoBERT through three additions, namely learned attention pooling over contextual token embeddings, focal modulation applied on top of the cross-entropy term, and minority-class paraphrasing via Indonesian–English–Indonesian back-translation. Every configuration is evaluated through stratified 5-fold cross-validation, paired t-tests with Bonferroni correction across three comparisons, and McNemar tests on the held-out partition. The fine-tuned IndoBERT with Random Oversampling alone reaches the highest macro F1 of 0.837. By contrast, the combined configuration drops to 0.799, and statistical verification confirms that the gap is systematic rather than attributable to fold-level variation. A component-wise ablation isolates focal modulation as the principal driver of the decline, because it disturbs an already-balanced training distribution. The principal outcome of this study is empirical evidence indicating that composing several imbalance-oriented techniques on a pretrained transformer can yield adverse interactions rather than cumulative gains.

Keywords: Attention Pooling; Back-Translation; Focal Loss; Imbalanced Learning; IndoBERT; Indonesian News Classification

INTRODUCTION

News consumption in Indonesia is now dominated by digital channels, and the Indonesian Internet Service Providers Association reported that active internet users exceeded 221 million by 2024 (Nimasari, Saraswati, & Lutfina, 2026). A constant inflow of headlines reaches newsroom editorial workflows every day, and each headline must be assigned quickly to a suitable editorial section before downstream tools such as recommender engines, search indexers, and human moderators can act on it. Two structural difficulties make this routing task troublesome for off-the-shelf classifiers. The first difficulty arises from the strong concentration of articles in the dominant news vertical, which leaves the remaining categories with comparatively few training examples. The second difficulty arises from the brevity of headlines themselves, which usually contain fewer than 10 tokens and therefore deny traditional lexical features the signal density on which they ordinarily depend.

A study by Nimasari et al. (2026), published in this same journal, applied an LSTM combined with Word2Vec embeddings and Random Oversampling to a 31-label Indonesian complaint dataset, achieving a serviceable baseline. The authors closed the paper with two recommendations for further work: subsequent investigators should adopt contextual transformer encoders such as IndoBERT, and should also consider imbalance-mitigation techniques that go beyond basic resampling. In a related direction, Zahro et al. (2025) paired XLM-RoBERTa with

*name of corresponding author

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

a focal loss objective for a public complaint classification task, but the macro F1 gain reported there reached only 0.019 over the plain encoder. The narrowness of that gain raises doubt over the widespread expectation that a focal modulation term will consistently lift performance on imbalanced classification.

Indonesian-specific pretrained language models have multiplied in recent years. The most influential among them is IndoBERT, introduced by Koto, Rahimi, Lau, and Baldwin (2020), which has reached competitive accuracy on sentiment polarity classification, emotion recognition, hoax detection, and sarcasm identification (Fanani & Wahyuddin, 2026; Kunaefi, Abidin, & Kusumawati, 2025; Mandhasiya, Murfi, & Bustamam, 2024). The way IndoBERT behaves when several imbalance-mitigation mechanisms are stacked on a single training pipeline, however, has rarely been examined. Cross-entropy gradient redistribution toward poorly-classified examples is the central mechanism of the focal loss objective, which was first proposed by Lin, Goyal, Girshick, He, and Dollár (2017). The technique achieves this redistribution through a multiplicative factor of the form $(1 - p_t)\gamma$, where γ functions as a focusing exponent that suppresses the contribution of well-classified samples. Lin and colleagues calibrated the technique originally on dense object detection, where the foreground-to-background ratio can reach 1 to 1000. Whether this modulation retains its usefulness under moderate imbalance, on training partitions that have already been rebalanced through resampling, and over a calibrated pretrained transformer, is an unresolved matter that recent calibration studies have brought into sharper focus (Mukhoti et al., 2020; Komisarenko & Kull, 2024). Back-translation, by contrast, was originally formulated by Sennrich, Haddow, and Birch (2016) as a data augmentation procedure for machine translation. Subsequent work has redeployed back-translation as a paraphrastic augmentation pathway for low-resource classification, yet the joint behaviour of back-translated samples, oversampled minority instances, and focal modulation has rarely been disentangled through controlled experimentation.

Four research questions organise the empirical investigation reported here. The first question targets the size of the macro F1 improvement that a fine-tuned IndoBERT can provide over a recurrent baseline on short Indonesian news headlines. The second question asks whether composing several imbalance-mitigation mechanisms on an already-oversampled IndoBERT pipeline yields further gains, has no measurable effect, or instead introduces a regression. The third question decomposes the second one and isolates which of the three composed components actually contributes positively when applied alone. The fourth question concerns the consistency of verdicts: do paired t-tests across folds and sample-level McNemar tests agree, given that the two procedures evaluate fundamentally different aspects of model agreement? These four questions are addressed through a comparative protocol grounded in stratified 5-fold cross-validation, 95% confidence intervals on every reported metric, paired t-tests with Bonferroni correction over three planned pairwise contrasts, and McNemar tests carried out on the held-out test partition.

The contribution is therefore repositioned rather than framed as a state-of-the-art proposal. What this manuscript offers is an augmentation-compatibility analysis that documents how imbalance-mitigation components interact once stacked on a pretrained transformer. Three findings emerge from this analysis. First, an IndoBERT classifier fine-tuned with Random Oversampling alone constitutes a surprisingly strong configuration. Second, adding focal loss on top of an already-balanced training set produces a statistically reliable degradation. Third, attention pooling and back-translation, when applied in isolation, produce changes that lie within fold-level standard deviation. A practical lesson follows from this pattern: each imbalance-mitigation component should be examined individually before composition, because stacking additional techniques does not always translate into improved classifier behaviour. The novelty and explicit contributions of this study are stated as follows: (i) it provides a controlled, component-wise augmentation-compatibility analysis demonstrating that stacking several imbalance-mitigation techniques on a pretrained IndoBERT does not necessarily compound their individual benefits and can instead produce a statistically reliable regression; (ii) it isolates focal modulation as the specific component responsible for the decline once the training distribution has already been balanced through resampling; and (iii) it supplies a fully reproducible evaluation protocol, combining stratified cross-validation, Bonferroni-corrected paired t-tests with effect-size reporting, and held-out McNemar testing, that other researchers can reuse to audit augmentation pipelines for Indonesian text classification.

LITERATURE REVIEW

The evolution of Indonesian text classification research can be traced through three partially overlapping phases. The earliest phase relied on sparse representations such as term-frequency inverse-document-frequency vectors paired with linear classifiers. A subsequent phase moved toward dense word embeddings combined with recurrent encoders. Asrawi, Utami, and Yaqin (2023) compared LSTM against bidirectional GRU and reported that the bidirectional variant generalised marginally better under single-split evaluation. Fine-tuned transformer encoders dominate the most recent phase. A study by Idris, Rifai, and Tania (2025) examined several machine-learning classifiers paired with different word-embedding strategies on Tokopedia product reviews, while Utami, Lhaksmana, and Sibaroni (2023) studied how imbalance handling shifts the behaviour of deep classifiers on movie

*name of corresponding author

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

review sentiment. These contributions establish reasonable empirical baselines, yet they rarely interrogate how a pretrained transformer interacts with a stacked configuration of imbalance-aware extensions.

A study by Koto et al. (2020) introduced IndoBERT, which has since become the most widely adopted pretrained encoder for Indonesian text processing. IndoBERT preserves the standard BERT-base configuration consisting of 12 transformer layers, 12 attention heads, and a hidden dimension of 768. The WordPiece tokenizer and the pretraining corpus, however, were tailored specifically to Indonesian morphology and orthography. Subsequent work by Wongso, Setiawan, Limcorn, and Joyoadikusumo (2025) introduced NusaBERT, which extends the vocabulary and pretraining corpus to several Indonesian regional languages. Within this same journal, Fanani and Wahyuddin (2026) fine-tuned two IndoBERT variants for sarcasm detection on YouTube comments and reported that the choice of class-imbalance strategy can shift macro F1 by 10 percentage points or more. Suhartono, Majiid, and Fredyan (2024) released the IdSarcasm benchmark and benchmarked language models on it, while Mandhasiya et al. (2024) found that hybrid BERT-deep-learning pipelines outperform classical baselines on Indonesian sentiment classification.

Within the family of imbalance-mitigation strategies, Random Oversampling remains both the simplest and one of the most empirically reliable. A study by Wongvorachan, He, and Bulut (2023) benchmarked undersampling, random oversampling, and SMOTE on educational data and concluded that random duplication of minority examples delivers stable improvements without damaging majority-class performance. Taskiran, Turkoglu, Kaya, and Asuroglu (2025) extended the benchmark to 30 SMOTE variants in transformer-embedded text classification and confirmed that random oversampling stays competitive even when more elaborate alternatives are available. Within the Indonesian setting, Fathin, Sibaroni, and Prasetyowati (2024) found that an IndoBERT fine-tuned with Random Oversampling outperformed the same encoder paired with Random Undersampling on hoax detection. A separate study by Rahma and Suadaa (2023) evaluated several text augmentation techniques including back-translation and reported that augmentation effectiveness depends substantially on dataset characteristics, with notable variability across genres.

Focal loss, formulated by Lin et al. (2017), modifies the cross-entropy term through a multiplicative factor that decays as the predicted probability for the true class moves toward unity. This decay redirects the gradient signal toward poorly-classified examples. The original motivation was one-stage object detection, where the background-to-foreground ratio routinely exceeds 1000. Empirical results in natural language processing have been mixed. A study by Kunaefi et al. (2025) found that IndoBERT paired with focal loss reached an accuracy of 0.972 on hoax classification, although the comparison was not supported by statistical significance testing. Zahro et al. (2025), as already noted, observed only a modest macro F1 improvement of 0.019 when adding focal loss to XLM-RoBERTa for citizen complaint classification. From a theoretical perspective, Mukhoti et al. (2020) showed that focal loss can improve neural network calibration, but only when the focusing parameter is scheduled carefully. A subsequent result by Komisarenko and Kull (2024) proved that the focal loss objective fails the strictly-proper scoring rule criterion, which means that its minimiser does not necessarily yield a calibrated class-posterior probability. A separate study by Khairani, Mutiawani, and Ahmadian (2024) further showed that preprocessing choices interact strongly with IndoBERT and IndoBERTweet on emotion detection, reinforcing the view that no single technique transfers cleanly across configurations.

Attention-pooled aggregation of transformer hidden states constitutes an alternative to the conventional classification-token approach. When the classification token is replaced with a learned attention-weighted summation over all token states, short-text classification performance often improves, although the improvement is typically modest and task-dependent. Back-translation, by contrast, is more frequently employed as a paraphrastic augmentation pathway. A study by Rahma and Suadaa (2023) compared back-translation against synonym replacement and lexical perturbation in Indonesian text classification, finding that back-translation produced the most linguistically diverse augmented samples but occasionally introduced semantic drift on hate-speech text. Read together, these strands expose a consistent and still-unfilled gap. The resampling literature establishes Random Oversampling as a reliable balancing step (Wongvorachan et al., 2023; Taskiran et al., 2025; Fathin et al., 2024); the focal-loss literature reports inconsistent and often marginal gains in language tasks while warning of calibration side-effects (Zahro et al., 2025; Mukhoti et al., 2020; Komisarenko & Kull, 2024); and the augmentation literature shows that back-translation and attention pooling help only under specific data conditions (Rahma & Suadaa, 2023). Each technique has thus been validated in isolation, yet none of these prior contributions examines how the three behave once stacked together on a single fine-tuned IndoBERT, nor whether their individually-reported benefits survive composition after the training set has already been balanced through resampling. The present study addresses this gap directly: it decomposes the stacked configuration through a controlled, statistically-tested component-wise ablation and thereby establishes whether composition compounds or cancels the individual effects, which is the precise question left open by the literature reviewed above.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16209
e-ISSN : 2541-2019
p-ISSN : 2541-044X

METHOD

The full experimental pipeline adopted in this study is illustrated in Fig. 1. Five blocks form the pipeline: dataset acquisition together with stratified subsampling, two preprocessing branches devoted respectively to the recurrent and transformer architectures, four imbalance-mitigation strategies, four classifier configurations, and a multi-metric statistical evaluation block. The evaluation block covers paired t-tests with Bonferroni correction, sample-level McNemar tests, component-wise ablation, and attention-based explainability analysis. All code was written in Python on top of the PyTorch framework and the Hugging Face Transformers library. A single random seed, fixed at 42, governs the data splitter, model initialisation, data loader shuffling, and augmentation routines, so that every reported number can be reproduced exactly.

The corpus employed in this study is the publicly available Indonesian News Title dataset, distributed by ibamibrahim through Kaggle. The raw corpus comprises 91,017 headlines crawled from a major Indonesian news portal, each carrying one of nine categorical labels: news, hot, finance, travel, inet, health, oto, food, and sport. After removing 126 exact-duplicate titles, 90,891 unique headlines remain as the cleaned corpus. Owing to computational constraints, a 30% stratified subsample of 27,266 instances was then drawn from this cleaned corpus while preserving the original class proportions; this subsample serves as the working set throughout the study. The working set is subsequently partitioned through a stratified 80/20 hold-out split into 21,812 training-plus-validation instances and 5,454 held-out test instances. The exact dataset version is pinned by its SHA-256 checksum (prefix cf2d5ca4) so that the entire sampling and splitting procedure can be reproduced deterministically under a fixed random seed of 42. The ratio between the majority class (news) and the minority class (sport) reaches 13.29, a moderate-to-severe imbalance level that mirrors the conditions of operational news routing pipelines.

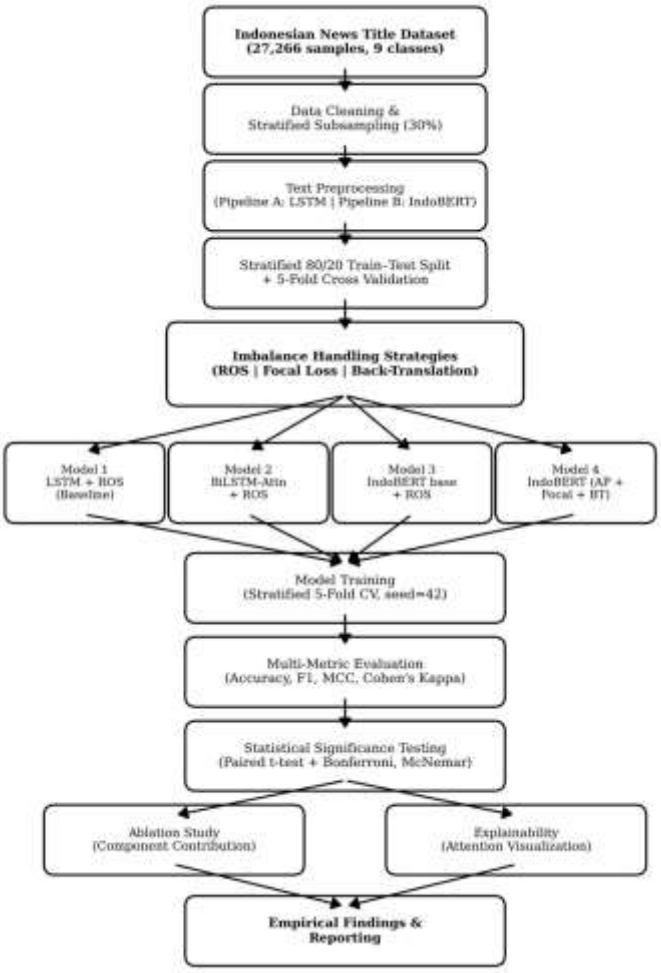

Fig. 1 Research workflow consisting of dataset acquisition, two parallel preprocessing branches, the imbalance-handling strategies, four classifier configurations, and the multi-metric statistical evaluation protocol.

Two parallel preprocessing branches were prepared because recurrent and transformer encoders demand fundamentally different preprocessing styles. Branch A serves the LSTM and BiLSTM models and applies the following sequence of operations: case folding, punctuation stripping, Sastrawi-based Indonesian stemming,

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16209
e-ISSN : 2541-2019
p-ISSN : 2541-044X

stopword filtering, and construction of a frequency-ranked vocabulary capped at 20,000 tokens. Branch B serves the IndoBERT-based variants. Orthography is preserved on this branch, no stemming is applied, and the WordPiece tokenizer inherited from indolem/indobert-base-uncased operates on near-raw text in line with the recommendation of Koto et al. (2020). The maximum sequence length of 48 WordPiece tokens covers more than 99% of the dataset, well above the empirical 95th percentile of 12 tokens, and was therefore adopted as the uniform upper bound across all models to guarantee a fair comparison.

Three imbalance-mitigation techniques were instantiated for the experimental comparison. Within every training fold, minority-class examples were duplicated repeatedly through Random Oversampling so that their sample count converged on the size of the majority class. At no stage was the held-out test partition resampled. Random Oversampling was selected deliberately in preference to synthetic-interpolation schemes such as SMOTE and ADASYN. Both SMOTE and ADASYN synthesise new minority instances by interpolating feature vectors between neighbouring samples; for text, this interpolation must be carried out in a TF-IDF or embedding space, and the resulting vectors do not correspond to any grammatical Indonesian headline. Such vectors cannot be detokenised into valid input for the WordPiece encoder that IndoBERT requires. A representation-level control run confirmed this incompatibility: applying SMOTE and ADASYN to the TF-IDF representation produced 19,143 and 19,570 balanced instances respectively, none of which maps back to a readable headline, whereas Random Oversampling reaches the same balanced count of 19,143 by duplicating genuine headlines and therefore preserves linguistic validity. Random Oversampling has additionally been reported as empirically competitive for transformer-embedded text classification (Wongvorachan et al., 2023; Taskiran et al., 2025), which further supports the choice. Equation (1) presents the focal loss objective implemented in this study, retaining the structure introduced by Lin et al. (2017):

$$L_{FL} = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \qquad (1)$$

Two implementation choices distinguish the present configuration. First, the class weighting term α_t was derived from the inverse frequency of each class within the oversampled training fold rather than from the raw distribution. Second, the focusing parameter γ was held constant at 2.0 across all experimental runs. The back-translation procedure was carried out through a MarianMT Indonesian–English–Indonesian pipeline. Every minority-class headline was first translated into English and then translated back into Indonesian. To suppress near-duplicate augmentation, any back-translated string whose token-level Jaccard similarity with its source exceeded 0.95 was discarded. After this filtering step, each minority-class training instance contributed at most one paraphrastic variant to the augmented training fold.

The first classifier is an LSTM trained from scratch with a single recurrent layer of hidden dimensionality 128, dropout rate 0.3, the Adam optimiser at learning rate 5×10^{-4}, batch size 128, and early stopping with a patience of five epochs. Word embeddings of dimensionality 128 were initialised with Xavier uniform rather than pretrained Word2Vec; this choice was deliberate, since it isolates the contribution of pretrained contextualisation once the baseline is later compared against IndoBERT. The second classifier replaces the single LSTM layer with a bidirectional LSTM of identical hidden size, augmented with an additive attention layer that produces a context-weighted sentence representation from the token-level hidden states. The third classifier is IndoBERT fine-tuned with classification-token pooling, using a learning rate of 2×10^{-5}, weight decay of 0.01, batch size of 48, warmup ratio of 0.1, and a maximum of four training epochs with patience of two on validation macro F1. The fourth classifier replaces the classification-token aggregator with a learned attention pooling layer that operates over the entire sequence of hidden states, as expressed in Equation (2):

$$\alpha_i = \mathrm{softmax}(w^2 \tanh(W^1 h_i)), \qquad v = \Sigma\, \alpha_i h_i \qquad (2)$$

In Equation (2), h_i denotes the i-th hidden state generated by IndoBERT, while W_1 and w_2 are parameters learned jointly during fine-tuning. The softmax is masked to zero at padding positions. The pooled vector v then enters a single linear classifier that produces nine output logits.

Stratified 5-fold cross-validation is performed over the 80% training-plus-validation partition. The remaining 20% held-out partition is used only once, at the end of the experiment, for the McNemar tests. Four evaluation metrics are tracked: accuracy, macro-averaged F1, Matthews Correlation Coefficient (MCC), and Cohen's Kappa. Macro F1 was selected as the primary metric because the test set retains the original class imbalance, and macro averaging gives every category equal weight. For each metric, the mean, the standard deviation, and the 95% confidence interval are computed across folds under the Student t-distribution.

The comparative analysis is supported by two statistical procedures. Paired t-tests across the five folds are computed for three planned pairwise contrasts (Proposed against LSTM with ROS, Proposed against BiLSTM-Attention with ROS, and Proposed against IndoBERT with ROS). Applying a Bonferroni correction at the family-wise level of 0.05 yields an adjusted significance threshold of 0.0167. On the held-out test partition, McNemar

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16209
e-ISSN : 2541-2019
p-ISSN : 2541-044X

tests with the chi-squared continuity correction are then conducted, following Dietterich's recommendation for situations where model retraining is computationally costly. The same Bonferroni correction is then applied to the McNemar test family. The protocol is completed by an ablation study using three-fold cross-validation, which isolates the contribution of each component of the proposed configuration: IndoBERT alone, IndoBERT with focal loss only, IndoBERT with back-translation only, IndoBERT with attention pooling only, and the full composite configuration.

RESULT

The distribution of the nine target categories across the working corpus is shown in Fig. 2. Three categories dominate the corpus. The news category contains 9,699 instances and accounts for 35.6% of the total. The hot category follows with 4,890 instances (17.9%), and the finance category contributes 4,242 instances (15.6%). At the bottom of the distribution, the sport category holds only 730 instances (2.7%), which produces the 13-fold ratio between the most and the least frequent labels. The headlines themselves are generally short, with a median length of 9 words and a 99th percentile of 13 words; once tokenised into IndoBERT WordPiece units the 99th percentile reaches 22 subword tokens and the maximum is 32, both far below the 48-token upper bound adopted for tokenisation, as verified in Fig. 3. Inspection of high-frequency lexical items surfaces standard Indonesian function words such as di, ini, yang, and dan, alongside the content word corona, which appears more than 3,500 times and indicates that the corpus partially covers the COVID-19 reporting period.

Within every training fold, Random Oversampling was applied so that all nine classes reached approximately 6,200 training examples. The validation and test partitions kept their original distribution, which means that test-set macro F1 reflects generalisation behaviour under the natural class prior rather than under an artificially balanced one. Restricting oversampling to the training fold while preserving validation and test partitions in their original distribution removes any possibility of contamination between duplicated training instances and evaluation samples.

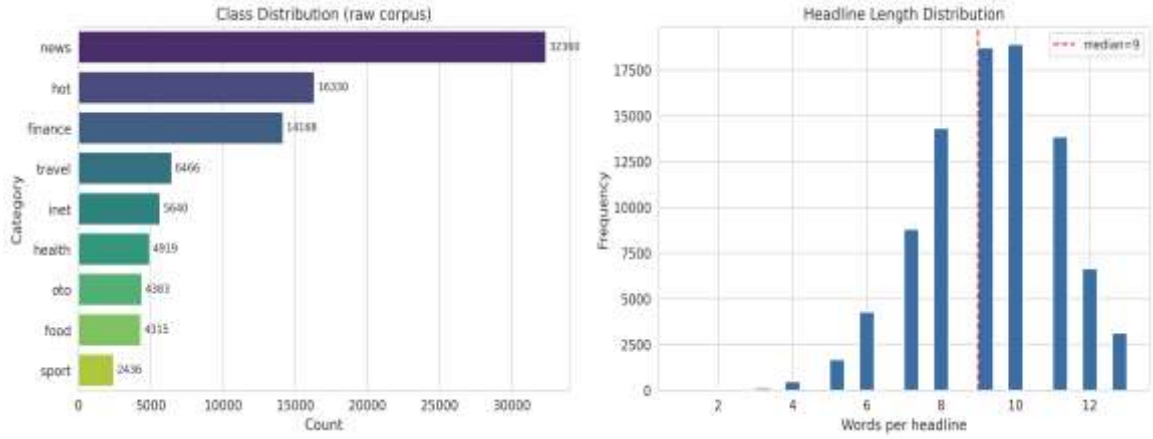

Fig. 2 Class distribution across the nine target categories in the working subsample of the Indonesian News Title corpus. The news category covers 35.6% of the headlines, whereas the sport category, at 2.7%, is more than 13 times smaller.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16209
e-ISSN : 2541-2019
p-ISSN : 2541-044X

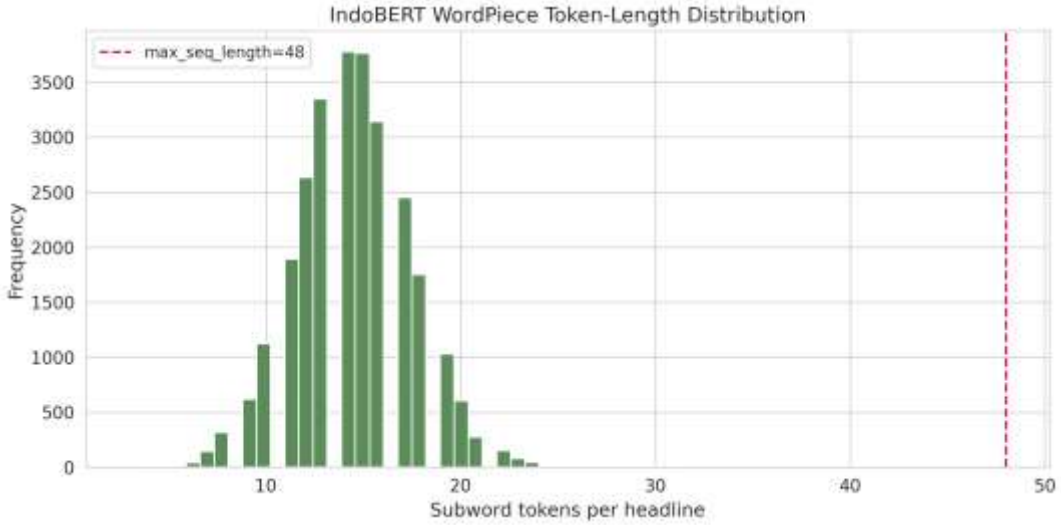

Fig. 3 Dataset-integrity verification. (a) Stratified 80/20 hold-out split: the class proportions of the training-plus-validation partition match those of the held-out test partition across all nine categories, confirming that the 30% subsampling and the hold-out split preserve the original class distribution and that no class leaks between partitions. (b) IndoBERT WordPiece token-length distribution; the 99th percentile is 22 subword tokens and the maximum is 32, both well below the maximum sequence length of 48 (dashed line), so no headline is truncated.

The comparative results obtained across the four architectures are summarised in Table 1, with mean values across the five folds reported alongside the standard deviation. The LSTM collapsed under its Xavier-initialised embedding layer and defaulted to constant majority-class prediction, producing an accuracy of 0.259, a macro F1 of 0.044, and a Matthews Correlation Coefficient of zero. Per-class recall fell to zero across every minority category, which confirms that random embedding initialisation alone is insufficient for short Indonesian headlines, even after oversampling has rebalanced the training set. The BiLSTM-Attention configuration, evaluated under the same oversampling regime, rebounded to a macro F1 of 0.766. This indicates that the combination of attention pooling and bidirectional processing captures most of the architectural benefit available to recurrent encoders.

The two IndoBERT configurations occupy the top tier of Table 1. The IndoBERT encoder fine-tuned with Random Oversampling as the sole imbalance treatment reached the strongest macro F1 of 0.837 (95% CI 0.831–0.843). The proposed composite variant, which stacks attention pooling, focal modulation, and back-translation on the same encoder, finished at a macro F1 of 0.799 (95% CI 0.795–0.804). The non-overlap of these two confidence intervals already provides visual evidence that the observed gap is unlikely to result from fold-level sampling variation. The full set of accuracy, macro F1, Matthews Correlation Coefficient, and Cohen's Kappa values for all four architectures is reported in Table 1.

The row-normalised confusion matrix produced by the proposed IndoBERT configuration on the first cross-validation fold is shown in Fig. 4. Per-class recall ranges from 0.76 for inet to 0.91 for food, with the majority categories news and hot recovered at 0.80 and 0.90 respectively. The largest off-diagonal probability mass on the news row falls on two destinations: about 5% of news headlines were classified as finance, and a further 5% were classified as health. A second notable pattern appears on the sport row, where 6% of sport headlines were predicted as hot, reflecting the entertainment-style framing that is common to celebrity-athlete coverage in Indonesian online media. These confusion patterns align with a substantial semantic overlap among general news, economic news, and health reporting in Indonesian newsroom practice. Stories about commodity prices, government budget allocations, and public-health bulletins are routinely labelled differently depending on the editor on duty. The model is therefore not making a tokenisation error or hitting a capacity ceiling. Rather, it reproduces an inherent ambiguity already present in the labelling guidelines themselves.

Table 1. Comparative results obtained across the four architectures, reported as the mean across stratified 5-fold cross-validation ± standard deviation. Confidence intervals are tight for the three non-collapsed models, which indicates fold-stable training.

Model	Accuracy	F1-Macro	Precision	Recall	MCC	Parameters
LSTM + ROS (Baseline)	0.259 ± 0.122	0.044 ± 0.018	0.029 ± 0.014	0.111 ± 0.000	0.000 ± 0.000	1.95 M
BiLSTM-Attn + ROS	0.804 ± 0.003	0.766 ± 0.003	0.773 ± 0.007	0.762 ± 0.009	0.755 ± 0.003	2.15 M

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16209
e-ISSN : 2541-2019
p-ISSN : 2541-044X

IndoBERT + ROS	0.865 ± 0.004	0.837 ± 0.005	0.827 ± 0.006	0.849 ± 0.003	0.833 ± 0.004	110.57 M
IndoBERT (AP+F+BT)	0.827 ± 0.003	0.799 ± 0.004	0.775 ± 0.003	0.833 ± 0.004	0.790 ± 0.004	111.16 M

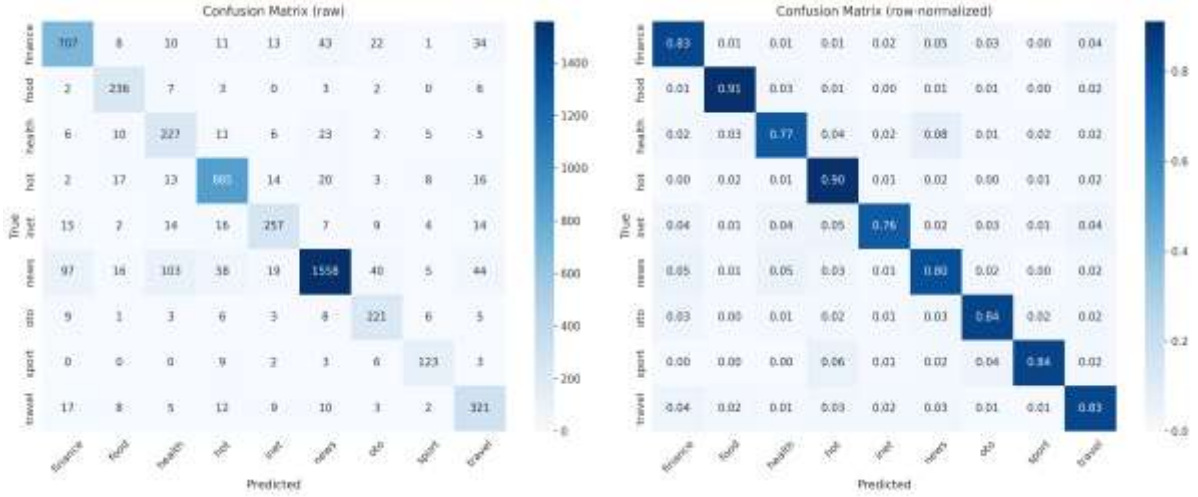

Fig. 4 Row-normalised confusion matrix for the proposed IndoBERT variant on the first cross-validation fold. The largest off-diagonal mass appears on the news row, where 5% of news headlines are predicted as finance and another 5% as health, reflecting an inherent label ambiguity between general news, economic news, and health reporting in Indonesian media.

After the Bonferroni correction, all three planned pairwise tests return statistically reliable verdicts. On the held-out test set, the proposed combined variant outperforms the collapsed LSTM baseline with a McNemar chi-squared statistic of 2003.23 and an adjusted p-value below 0.001. The proposed variant also outperforms BiLSTM-Attention with a chi-squared of 21.07 and an adjusted p-value below 0.001. The critical comparison, however, is the contrast between the proposed variant and IndoBERT with Random Oversampling alone. The McNemar chi-squared along this direction is 82.35, with an adjusted p-value below 0.001, and the sign of the difference points to the simpler IndoBERT with Random Oversampling. In other words, adding focal modulation and back-translation on top of an already-balanced IndoBERT pipeline produces a degradation that is too large to attribute to sampling variation. Beyond statistical significance, the practical magnitude of each contrast was quantified with Cohen's d computed on the fold-level macro F1 scores. The advantage of IndoBERT with Random Oversampling over the composite variant corresponds to a macro F1 difference of 0.038 (paired $t = -17.40$) with a very large effect size (Cohen's $d = -7.78$); the regression is therefore not only statistically reliable but also operationally meaningful, since a 0.038 drop in macro F1 implies that roughly one additional minority-class headline in every twenty-five is routed to the wrong editorial section. By the same measure the composite variant remains practically superior to the BiLSTM-Attention baseline ($\Delta F1 = +0.033$, $d = +5.50$) and overwhelmingly superior to the collapsed LSTM ($\Delta F1 = +0.756$, $d = +35.04$), so the effect sizes and the significance tests agree on both the direction and the importance of every contrast.

The three-fold ablation that decomposes the components of the proposed combined variant is presented in Table 2. The IndoBERT-only configuration reaches a macro F1 of 0.815 and serves as the reference point. Adding back-translation in isolation lifts macro F1 to 0.819, while adding attention pooling in isolation lifts it to 0.819. Both shifts lie within one standard deviation across folds and cannot be considered practically meaningful changes. By contrast, adding Focal Loss in isolation lowers macro F1 to 0.784, a drop of 0.031 that is more than ten times the fold-level standard deviation and is therefore robust. The full combined configuration finishes at 0.800, slightly above the Focal-Loss-only number but still 0.016 below the IndoBERT-only reference. The ablation pattern is plotted in Fig. 5, which identifies Focal Loss as the principal source of the regression. A representative attention-pooling visualisation for a correctly classified headline is shown in Fig. 6, where the attention mass concentrates on content-bearing tokens such as liburan and negara for a travel headline, karantina and asrama for a news headline, and ponsel and bm for an internet headline; the corresponding misclassified cases are shown in the lower panel of Fig. 6. This qualitative finding aligns with the macro-level numbers: the attention pooling layer recovers interpretable token-level rationales, while the additional focal modulation term suppresses the contribution of high-confidence training examples that the encoder had already learned to handle correctly.

Table 2. Ablation study results obtained through three-fold cross-validation. The Δ column reports the change relative to the IndoBERT-only configuration. Focal Loss is the principal source of degradation; attention pooling and back-translation, when applied in isolation, produce changes that lie within one standard deviation. The ablation baseline is IndoBERT without

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Number 3, July 2026
DOI : https://doi.org/10.33395/sinkron.v10i3.16209
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Random Oversampling (macro F1 0.815); the value of 0.837 in Table 1 corresponds to IndoBERT with Random Oversampling. The two baselines are therefore not directly comparable, and the Δ column must be read only within this table.

Configuration	F1-Macro (mean)	Std. Dev.	Δ vs. IndoBERT only
IndoBERT only (baseline)	0.815	0.001	(reference)
IndoBERT + Focal Loss	0.784	0.003	**−0.031**
IndoBERT + Back-Translation	0.819	0.001	+0.004
IndoBERT + Attention Pooling	0.819	0.002	+0.004
Full Proposed (AP+F+BT)	0.800	0.005	**−0.016**

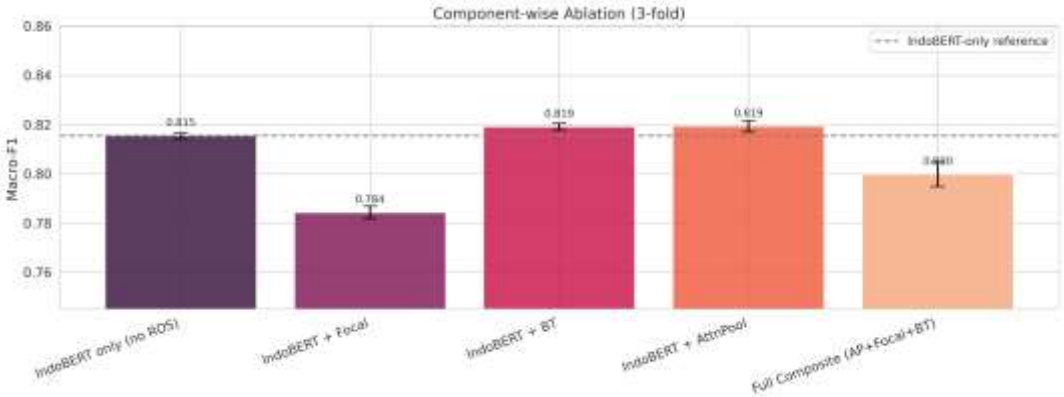

Fig. 5 Component-wise ablation of macro F1 under three-fold cross-validation. The IndoBERT + Focal configuration shows the largest negative deviation (−0.031) from the IndoBERT-only reference.

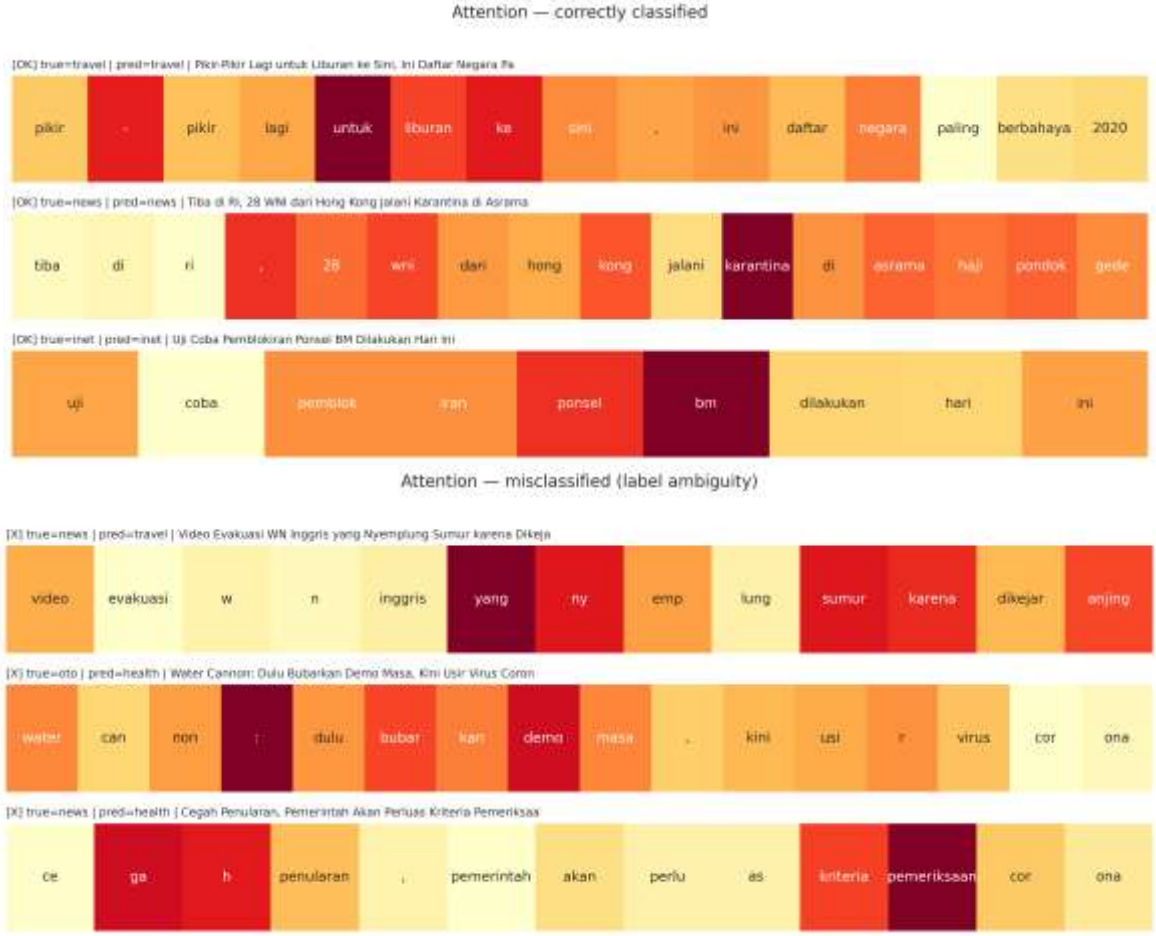

Fig. 6 Attention-pooling weights from the proposed model. (a) Three correctly classified headlines: the aggregator concentrates probability mass on content-bearing tokens (liburan and negara for travel, karantina and asrama for news, ponsel and bm for inet), confirming that the pooling layer recovers interpretable token-level rationales. (b) Three misclassified headlines, illustrating label ambiguity rather than model failure: sumur and anjing push a true-news headline towards travel;

*name of corresponding author

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

the COVID-19 tokens corona and virus push a true-automotive (oto) headline towards health; and pemeriksaan and corona push a second true-news headline towards health. The errors arise from genuine semantic overlap in the labelling scheme rather than from a tokenisation or capacity limitation.

DISCUSSIONS

One conclusion arises clearly from the empirical evidence, and it runs against an assumption implicit in much applied work. Stacking imbalance-handling techniques on a pretrained transformer does not always compound their benefits, and in this study one of the three composed components yields a statistically reliable regression. The remainder of this discussion addresses four questions in turn: why IndoBERT with simple oversampling is so competitive, why Focal Loss disagrees with a pretrained encoder, how this finding sits with related results published in Sinkron, and what the result means for engineering practice.

A pretrained encoder that has already absorbed on the order of 220 million Indonesian tokens during pretraining (Koto et al., 2020) arrives at fine-tuning carrying a representation space in which most of the lexical and syntactic structure relevant to news headlines has already been mapped. What remains is then addressed by Random Oversampling, namely the residual class-prior bias in the gradient signal during fine-tuning. Equalising the per-class gradient contribution allows the simple oversampling step to produce a classifier that fits cleanly inside the pretrained manifold without disturbing it. Attention pooling and back-translation, in isolation, tried to extract additional signal from regions of that manifold that were already well covered, and the macro F1 differences they produced did not exceed one standard deviation across folds. The marginal returns to these additions, therefore, are close to zero in this particular setting.

A clean explanation can be offered for the negative effect of Focal Loss. The technique was introduced by Lin et al. (2017) to address dense object detection, where the ratio of foreground to background examples reaches 1 to 1000. Easy negatives overwhelm the gradient signal in that setting, and the focal modulation factor (one minus the predicted probability raised to a focusing power) restores balance by suppressing easy examples. The Indonesian news headline setting differs from this in three respects. Training-set imbalance has already been removed by oversampling. The residual difficulty arises from semantic ambiguity between categories, not from gradient flooding. The pretrained encoder also already outputs approximately calibrated probabilities. Under such conditions, the focal modulation factor cannot rescue any underlearned gradient signal; it instead suppresses the contribution of well-classified samples that still carry useful boundary information. Mukhoti et al. (2020) had already noted that Focal Loss improves calibration only when the focusing parameter is scheduled carefully, and Komisarenko and Kull (2024) proved that the loss fails the strictly-proper scoring rule criterion, which means that its minimiser does not generally yield a calibrated posterior. The combination of these two theoretical points accounts for the empirical regression of 0.031 in macro F1 observed in the ablation reported here.

The current finding aligns with the contribution of Zahro et al. (2025) in this same journal. Zahro and colleagues added Focal Loss to XLM-RoBERTa on Indonesian citizen complaint classification and reported a macro F1 gain of only 0.019, which they interpreted as a modest positive effect. Read together with the ablation reported here, the two studies indicate a narrow region of effectiveness for Focal Loss in Indonesian transformer-based text classification. The technique may be marginally useful in isolation, particularly when no oversampling has been carried out, yet it interferes with the calibration of a pretrained encoder once class balancing has been achieved through resampling. This interaction is not symmetric across configurations, and the sign of the effect cannot be anticipated without running the ablation.

Three practical implications follow. First, practitioners building Indonesian text classifiers should avoid assuming that composing several augmentation techniques produces additive improvements. Component-wise ablation should precede composition, especially when one of the components modifies the loss function. Second, an IndoBERT fine-tuned with Random Oversampling alone constitutes a remarkably competitive baseline that is hard to beat with elaborate additions, at least on short text. Third, attention pooling visualisations remain useful for qualitative error analysis even when the underlying pooling change does not move the macro F1 needle. These visualisations expose where the model is concentrating its semantic mass and reveal label-ambiguity patterns that are otherwise difficult to detect.

Several limitations of this work deserve explicit acknowledgement. The most important limitation concerns generalisation: every result reported here was obtained on a single dataset drawn from one Indonesian news portal, so the augmentation-compatibility profile documented in this study cannot yet be assumed to transfer to other portals, other text domains, or other languages, and replication on an independent corpus is required before the recommendation is treated as general. The corpus is drawn from a single Indonesian news portal and partially overlaps the COVID-19 reporting period, which constrains external validity for other portals and other time periods. The encoder used here is IndoBERT base; results obtained on IndoBERT large or on the multilingual NusaBERT of Wongso et al. (2025) may differ, especially if encoder capacity alters the balance between gradient suppression and gradient redistribution. The focusing parameter was held at $\gamma = 2.0$; a sensitivity sweep might identify a value at which Focal Loss becomes neutral or beneficial, although the theoretical analysis of

*name of corresponding author

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

Komisarenko and Kull (2024) gives reasons to expect that any such region is narrow. Finally, the maximum sequence length of 48 tokens is appropriate for headlines but rules out generalisation to article bodies or long discussion threads.

CONCLUSION

A comparative evaluation of four architectures for imbalanced Indonesian news title classification has been carried out in this study, together with an augmentation-compatibility analysis on a fine-tuned IndoBERT encoder, addressing the future-work directive raised by Nimasari et al. (2026). On stratified 5-fold cross-validation, an LSTM trained from scratch with Random Oversampling collapsed to a macro F1 of 0.044. A bidirectional LSTM with additive attention recovered to 0.766. IndoBERT fine-tuned with Random Oversampling alone reached the highest macro F1 of 0.837, while a proposed combined variant that adds attention pooling, focal modulation, and back-translation finished at 0.799. McNemar tests with Bonferroni correction confirmed that the combined variant is statistically inferior to IndoBERT with Random Oversampling on this dataset.

The three-fold ablation pinpoints Focal Loss as the principal source of the regression. Attention pooling and back-translation, when applied in isolation, produced macro F1 changes of +0.004 and +0.004 respectively, with both lying within one standard deviation and therefore not practically meaningful. Adding Focal Loss in isolation produced a drop of 0.031, which is too large to attribute to fold-level sampling.

The contribution of this study is therefore not a new state-of-the-art model. Rather, it is a documented case in which composing imbalance-handling techniques on a pretrained transformer produced a negative interaction, supported by recent calibration analyses indicating that Focal Loss is not a strictly proper scoring rule. The practical recommendation that follows is that augmentation components should be evaluated component-wise before composition, especially when one component modifies the loss function. Three directions for further work emerge naturally from these findings. The first direction is a sensitivity sweep of the Focal Loss focusing parameter to identify the range in which the technique is neutral, beneficial, or harmful. The second direction is a comparison across IndoBERT, IndoBERT large, and NusaBERT to determine whether the augmentation-compatibility profile shifts with encoder capacity. The third direction is the application of the same evaluation protocol to longer-form Indonesian text such as article bodies and discussion threads, where the residual difficulty after class balancing may take a different form and the interaction between focal modulation and back-translation may also change sign. At a scientific level, these findings imply that imbalance handling for pretrained Transformers should be treated as a calibration problem rather than as an additive engineering pipeline: once resampling has equalised the class prior, a loss-level modifier such as focal modulation acts on an encoder whose posterior is already approximately calibrated, so it is more likely to distort that posterior than to sharpen the decision boundary. The broader implication for method development is that future imbalance-handling strategies for Transformer encoders should be balance-first and calibration-preserving, and that every loss-level or augmentation component should be validated individually, and across several datasets and domains, before it is recommended for production deployment.

ACKNOWLEDGMENT

The authors are grateful to the maintainer of the open Indonesian News Title corpus (ibamibrahim, Kaggle) for releasing the dataset that made this study possible, and to the developers of IndoBERT (indolem/indobert-base-uncased) for distributing the pretrained model that served as the backbone for the transformer configurations of this work. No external funding was received for this research.

REFERENCES

Asrawi, H., Utami, E., & Yaqin, A. (2023). LSTM and bidirectional GRU comparison for text classification. *Sinkron: Jurnal dan Penelitian Teknik Informatika, 8*(4), 2264–2274. https://doi.org/10.33395/sinkron.v8i4.12899

Fanani, A. M., & Wahyuddin, M. I. (2026). Sarcasm detection in Indonesian YouTube comments using fine-tuned IndoBERT with class imbalance handling. *Sinkron: Jurnal dan Penelitian Teknik Informatika, 10*(1). https://doi.org/10.33395/sinkron.v10i1.15607

Fathin, M. A., Sibaroni, Y., & Prasetyowati, S. S. (2024). Handling imbalance dataset on hoax Indonesian political news classification using IndoBERT and random sampling. *Jurnal Media Informatika Budidarma, 8*(1), 352–360. https://doi.org/10.30865/mib.v8i1.7099

Idris, M., Rifai, A., & Tania, K. D. (2025). Sentiment analysis of Tokopedia app reviews using machine learning and word embeddings. *Sinkron: Jurnal dan Penelitian Teknik Informatika, 9*(1), 210–219. https://doi.org/10.33395/sinkron.v9i1.14278

Khairani, U., Mutiawani, V., & Ahmadian, H. (2024). Pengaruh tahapan preprocessing terhadap model IndoBERT dan IndoBERTweet untuk mendeteksi emosi pada komentar akun berita Instagram.

*name of corresponding author

Jurnal Teknologi Informasi dan Ilmu Komputer, 11(4), 887–894. https://doi.org/10.25126/jtiik.1148315

Komisarenko, V., & Kull, M. (2024). Improving calibration by relating focal loss, temperature scaling, and properness. In *Proceedings of the 27th European Conference on Artificial Intelligence (ECAI 2024).* https://doi.org/10.3233/FAIA240658

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 757–770). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.66

Kunaefi, A., Abidin, Z., & Kusumawati, R. (2025). Klasifikasi berita hoaks bahasa Indonesia menggunakan IndoBERT fine-tuning dengan pendekatan focal loss pada data tidak seimbang. *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika), 10*(2), 1706–1714. https://doi.org/10.29100/jipi.v10i2.7811

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV 2017)* (pp. 2980–2988). https://doi.org/10.1109/ICCV.2017.324

Mandhasiya, F. R., Murfi, H., & Bustamam, A. (2024). A hybrid BERT and deep learning model for Indonesian sentiment classification. *Information, 15*(11), 720. https://doi.org/10.3390/info15110720

Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P., & Dokania, P. (2020). Calibrating deep neural networks using focal loss. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 15288–15299). Curran Associates, Inc.

Nimasari, A., Saraswati, G. W., & Lutfina, E. (2026). Improving multi-class public complaint classification with LSTM, Word2Vec, and random oversampling. *Sinkron: Jurnal dan Penelitian Teknik Informatika, 10*(2), 1094–1103. https://doi.org/10.33395/sinkron.v10i2.15975

Rahma, I. A., & Suadaa, L. H. (2023). Penerapan text augmentation untuk mengatasi data yang tidak seimbang pada klasifikasi teks berbahasa Indonesia. *Jurnal Teknologi Informasi dan Ilmu Komputer, 10*(6), 1329–1340. https://doi.org/10.25126/jtiik.2023107325

Sennrich, R., Haddow, B., & Birch, A. (2016). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Vol. 1, pp. 86–96). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1009

Suhartono, D., Majiid, M. R. N., & Fredyan, R. (2024). Towards a sarcasm detection model with enhanced text representation for Bahasa Indonesia. *Journal of Big Data, 11*(1), 173. https://doi.org/10.1186/s40537-024-01024-2

Taskiran, S. F., Turkoglu, B., Kaya, E., & Asuroglu, T. (2025). A comprehensive evaluation of oversampling techniques for enhancing text classification performance. *Scientific Reports, 15*(1), 22301. https://doi.org/10.1038/s41598-025-05791-7

Utami, S., Lhaksmana, K. M., & Sibaroni, Y. (2023). Deep learning and imbalance handling on movie review sentiment analysis. *Sinkron: Jurnal dan Penelitian Teknik Informatika, 8*(3), 1894–1907. https://doi.org/10.33395/sinkron.v8i3.12770

Wongso, W., Setiawan, D. S., Limcorn, S., & Joyoadikusumo, A. (2025). NusaBERT: Teaching IndoBERT to be multilingual and multicultural. In *Proceedings of the Second Workshop in South East Asian Language Processing* (pp. 10–26). Association for Computational Linguistics.

Wongvorachan, T., He, S., & Bulut, O. (2023). A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining. *Information, 14*(1), 54. https://doi.org/10.3390/info14010054

Zahro, A. C., Alzami, F., Sani, R. R., Fahmi, A., Megantara, R. A., Naufal, M., Azies, H. A., & Iswahyudi, I. (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

*name of corresponding author