



Indonesian Hate Speech Detection under Class Imbalance Using a Soft-Voting Ensemble of IndoBERTweet and IndoRoBERTa

Muhammad Alkaff¹, Eka Setya Wijaya², Fadliyanur³, Muhammad Bahit⁴, Sinar Nadhif Ilyasa⁵

^{1,2,3,4}Department of Information Technology, Faculty of Engineering, Universitas Lambung Mangkurat, Banjarmasin, Indonesia

⁵Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia

*m.alkaff@ulm.ac.id

Abstract

Hate speech detection on Indonesian social media remains challenging due to the coexistence of formal and highly colloquial language, as well as the moderate class imbalance typical of real-world datasets. Models trained under these conditions often skew toward the majority class and generalize poorly across linguistic registers. This study investigates whether a simple, training-free model-level ensemble can improve Indonesian hate speech detection under such conditions without resampling the data. IndoBERTweet and IndoRoBERTa, pretrained respectively on informal Twitter text and broader formal corpora, serve as complementary base models, and their class probabilities are combined through equal-weight soft voting. On the Indonesian Hate Speech Superset ($N = 14,306$), evaluated across five random seeds with paired significance testing, the soft-voting ensemble attains a macro-averaged F1 of 0.898 ± 0.003 and a macro recall of 0.899 ± 0.003 . It significantly outperforms a TF-IDF SVM baseline and the IndoRoBERTa base model, while showing no significant difference from the stronger IndoBERTweet base model and a trained logistic-regression stacking ensemble. Notably, the ensemble matches the stacking ensemble without any additional training stage or meta-learner, and a calibration analysis shows it improves probability calibration over both base models. These results indicate that equal-weight probability averaging is a simple, reproducible, and competitive strategy for Indonesian hate speech detection under moderate class imbalance.

Keywords: hate speech detection; ensemble learning; imbalanced data; indonesian text; transformer models

How to Cite: M. Alkaff, E. S. Wijaya, F. Fadliyanur, M. Bahit, and S. N. Ilyasa, "Indonesian Hate Speech Detection under Class Imbalance Using a Soft-Voting Ensemble of IndoBERTweet and IndoRoBERTa", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 10, no. 3, pp. 814 - 824, Jun. 2026.

Permalink/DOI: <https://doi.org/10.29207/resti.v10i3.7681>

Received: April 2, 2026

Accepted: June 15, 2026

Available Online: June 24, 2026

This is an open-access article under the [CC BY 4.0 License](#)

Published by [Ikatan Ahli Informatika Indonesia](#)

1. Introduction

Hate speech on social media has emerged as a significant challenge in online communication ecosystems, as it can amplify social polarization, discrimination, and harmful discourse [1-4]. The rapid expansion of user-generated textual content across social media platforms has made large-scale manual moderation impractical. This development necessitates automated detection systems capable of operating in real time and at high volume [5]. Within the field of Natural Language Processing (NLP), hate speech detection has consequently evolved into a critical text classification task, underpinning applications such as content moderation, platform governance, and online safety monitoring [6-8]. Despite substantial advances in contextual language modeling and deep learning-based

classification, accurately identifying hate speech in real-world social media environments remains technically challenging, particularly due to linguistic variability, contextual ambiguity, and naturally skewed class distributions [9], [10].

In the Indonesian context, hate speech detection involves additional linguistic and sociotechnical complexities beyond standard text classification settings. Indonesian social media discourse frequently blends formal language with highly informal expressions, including slang, abbreviations, creative orthographic variations, phonetic spellings, and culturally or regionally specific references. Moreover, code-mixing between Indonesian and other languages, particularly English and local dialects, together with non-standard grammatical constructions, increases

lexical variability and contextual ambiguity [11]. These characteristics challenge models that rely heavily on surface-level lexical features, since similar semantic content may appear in multiple orthographic forms. At the same time, subtle contextual cues often determine whether an expression is hateful. Consequently, effective hate speech detection in Indonesian social media requires models capable of learning robust contextual representations across heterogeneous linguistic registers [12], [13].

Beyond linguistic variability, class imbalance introduces additional methodological challenges in hate speech detection. In many real-world social media datasets, non-hateful content naturally exceeds hateful instances, reflecting the distribution of online discourse [14], [15]. Models trained under such skewed conditions may develop a bias toward the majority class, achieving high overall accuracy while exhibiting comparatively lower recall for hate speech [16]. This imbalance is particularly concerning in safety-sensitive applications, where misclassifying harmful content may undermine the effectiveness of automated moderation systems. For this reason, prior research has emphasized the use of imbalance-aware evaluation metrics, such as macro-averaged F1-score, which provide a more balanced assessment of performance across classes under uneven distributions [17]. These considerations highlight the importance of designing modeling strategies that maintain robustness without artificially altering the underlying data distribution.

Class imbalance in text classification is commonly addressed at either the data level or the algorithm level. Data-level methods, such as random oversampling, SMOTE, and ADASYN, rebalance the training distribution by duplicating or synthesizing minority-class instances; however, synthetic interpolation is poorly defined in high-dimensional contextual embedding spaces, and such methods have been reported to yield limited or inconsistent gains for deep NLP models unless interpolation is performed in a learned embedding space [18], [19]. Algorithm-level alternatives instead reshape the training objective to emphasize the minority class, including class-weighted cross-entropy, class-balanced loss based on the effective number of samples [20], and focal loss, which down-weights easy examples so that learning concentrates on difficult, typically minority-class, cases [21]. A related concern is probability calibration: training under imbalance can leave models overconfident, motivating post-hoc corrections such as temperature scaling [22]. In contrast to these interventions, the present study deliberately avoids data resampling and instead asks whether a model-level ensemble of two complementary pretrained encoders can provide robust performance under moderate imbalance without modifying either the data distribution or the loss function.

Early studies on hate speech detection largely relied on traditional machine learning algorithms, such as

Support Vector Machines, Naïve Bayes, and logistic regression, typically using bag-of-words or TF-IDF features [23]. While these approaches established important baseline benchmarks, their reliance on surface-level lexical representations limited their ability to capture contextual meaning, particularly in informal, highly variable social media text. Subsequent work introduced neural network-based architectures, including convolutional and recurrent models such as LSTM and BiLSTM, which improved performance by modeling sequential dependencies and distributed representations [24], [25].

In the context of Indonesian social media, however, empirical studies indicate that these architectures remain sensitive to lexical variation and class imbalance. Performance often deteriorates when minority-class instances are underrepresented or when class imbalance is addressed through synthetic sampling techniques [11], [12], [14]. These limitations highlight the need for more expressive contextual modeling approaches that can better generalize across diverse linguistic forms.

The Indonesian Hate Speech Superset dataset was introduced in [26] to mitigate issues of dataset fragmentation and inconsistent evaluation protocols in prior Indonesian hate speech research [27-29]. By consolidating multiple publicly available corpora into a unified benchmark and preserving their naturally occurring class distributions, the dataset provides a standardized, comparatively realistic evaluation setting. Its adoption in recent studies has facilitated more consistent cross-model comparisons [30]. Importantly, preserving the original class proportions makes this dataset particularly suitable for examining modeling strategies that aim to improve robustness under moderate class imbalance without relying on synthetic resampling.

Recent advances in NLP largely stem from Transformer-based architectures, which learn contextual representations through large-scale pretraining and consistently outperform earlier neural models in text classification tasks [31], [32]. In the Indonesian setting, pretrained models such as IndoBERTweet and IndoRoBERTa have demonstrated competitive performance across various social media and language understanding benchmarks [33], [34]. IndoBERTweet is pretrained on Indonesian Twitter data, making it well-suited for modeling informal, slang-rich, and highly variable social media language [35]. In contrast, IndoRoBERTa is pretrained on a broader range of Indonesian texts, including more formal sources, which helps it capture representations aligned with standardized linguistic structures [36]. Due to differences in their pretraining data and exposure to various domains, combining their predictions may provide complementary contextual signals that enhance robustness across heterogeneous linguistic inputs.

Ensemble learning has been investigated as a model-level strategy to improve robustness and generalization

in text classification and hate speech detection tasks [37], [38]. By aggregating predictions from multiple models, ensemble approaches can reduce prediction variance and partially compensate for individual model biases [39]. Existing ensemble methods in NLP often employ stacking or multi-stage training schemes, which introduce additional architectural complexity and computational overhead [40], [41]. While such strategies may yield performance gains, their benefits in real-world class-imbalanced settings remain context-dependent [42]. In contrast, probability-level aggregation methods, such as soft voting, provide a simpler alternative that integrates model outputs without altering the original data distribution or requiring additional training procedures [43], [44]. However, prior work has not systematically examined the effectiveness of such lightweight ensemble strategies for Indonesian hate speech detection under naturally imbalanced conditions.

In light of these considerations, this study investigates whether a simple, training-free, probability-level ensemble can perform competitively with stronger modeling strategies for Indonesian hate speech detection under moderate, naturally occurring class imbalance, without modifying the original data distribution. The contributions of this work are threefold. First, we show that equal-weight soft voting over IndoBERTweet and IndoRoBERTa attains performance statistically comparable to both the strongest individual Transformer model and a trained logistic-regression stacking ensemble, while significantly surpassing a TF-IDF SVM baseline and the less competitive IndoRoBERTa base model, all without resampling, class re-weighting, or an additional meta-learner training stage. Second, we provide a rigorous empirical evaluation on a consolidated Indonesian benchmark, reporting results across five random seeds with paired significance testing using bootstrap and McNemar's tests, together with probability-calibration analysis; we also report comparable or non-significant differences transparently rather than overstating performance gains. Third, we present a qualitative error taxonomy of false negatives and false positives that characterizes persistent linguistic challenges for current models, including implicit political framing, identity slurs, code-mixing, and annotation noise. Together, these contributions position equal-weight soft voting as a simple, reproducible, and practical ensemble strategy for Indonesian hate speech detection under moderate class imbalance.

The remainder of this paper is organized as follows. Section 2 describes the dataset, the experimental design, the model fine-tuning procedure, and the construction of the soft-voting ensemble. Section 3 presents the experimental results, including multi-seed performance evaluation, statistical significance testing, probability-calibration analysis, and qualitative error analysis. Finally, Section 4 concludes the paper by summarizing

the main findings, discussing limitations, and outlining directions for future work.

2. Methods

This study employs a systematic experimental design to investigate ensemble learning strategies for Indonesian hate speech detection under imbalanced data conditions. The methodology focuses on integrating pretrained Transformer-based models via a probability-level soft-voting ensemble and comparing this approach with individual base models and a stacking-based ensemble. The experimental design comprises dataset preparation and stratified data splitting, fine-tuning of base Transformer models, ensemble construction, baseline comparison, and performance evaluation using metrics appropriate for imbalanced classification. All experiments are conducted using a fixed stratified data split and multiple training seeds to ensure reproducibility, and evaluation is performed exclusively on a held-out test set. The overall experimental pipeline is shown in Figure 1.

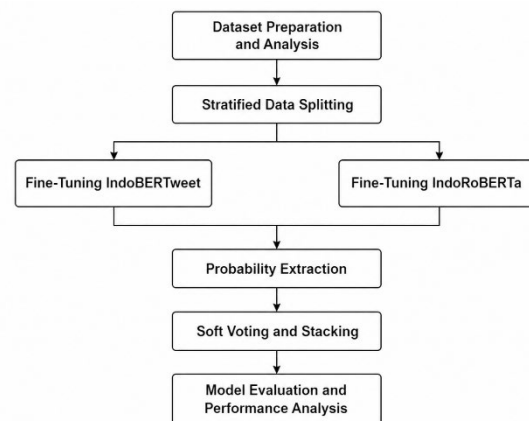


Figure 1. Experimental pipeline of the proposed ensemble approach.

2.1 Dataset and Data Splitting Strategy

This study uses the Indonesian Hate Speech Superset [26], a consolidated benchmark comprising 14,306 Indonesian social media posts from publicly available, documented datasets identified in a 2024 systematic survey. All labels are binarized into hate speech (1) and non-hate content (0).

The dataset aggregates three source corpora: IDHSD, the Dataset for Hate Speech Detection in Indonesian [29], which contains Twitter posts with binary hate/non-hate labels; ID_multilabel [28], which consists of Indonesian tweets originally annotated for hate speech, abusive language, and fine-grained target and severity categories; and ID_instagram [27], which contains Indonesian Instagram comments. During dataset construction, duplicate posts were removed, and usernames and URLs were replaced with fixed placeholder tokens to protect user privacy. In the current experimental split, the combined dataset shows moderate class imbalance: 57.7% non-hate instances and 42.3% hate instances.

To support ensemble learning and ensure a fair, unbiased evaluation, the dataset is divided into three disjoint subsets via stratified sampling to preserve the original class distribution across all splits. Specifically, 20% of the data is reserved as a held-out test set and excluded from all training and model selection stages. The remaining 80% is further split into a base-model training subset (60% of the total data) and a meta-training validation subset (20% of the total data). The base-model training subset is used to fine-tune the Transformer-based classifiers. In contrast, the meta-training validation subset is used to generate out-of-sample predictions for training the stacking ensemble's meta-learner. The held-out test set is used exclusively for final performance evaluation.

This three-way data-splitting strategy is adopted to prevent information leakage between the training, validation, and evaluation phases, which is particularly important in stacking-based ensemble settings. All splits are performed using a fixed random seed and stratification by class label to ensure reproducibility and consistent class proportions across subsets. The distribution of samples across the training, validation, and test subsets, including the corresponding class ratios, is summarized in Table 1.

Table 1. Dataset distribution across training, validation, and test sets

Subset	Total	Non-Hate	Hate	Hate Ratio (%)
Train	8583	4953	3630	42.29%
Validation	2861	1651	1210	42.29%
Test	2862	1652	1210	42.28%

2.2 Base Transformer Models and Fine-Tuning

Two pretrained Transformer-based models are employed as base classifiers: IndoBERTweet and IndoRoBERTa. These models are selected because they represent complementary Indonesian-language encoders and have demonstrated competitive performance on Indonesian text classification tasks.

IndoBERTweet is a BERT-based encoder pretrained on Indonesian Twitter data. It was developed through domain-adaptive pretraining of a general-purpose Indonesian BERT model on approximately 409 million word tokens collected from 26 million Indonesian tweets, using a domain-specific lowercased WordPiece vocabulary of 31,984 tokens. Its Twitter-domain pretraining makes it particularly suitable for modeling informal, abbreviation-heavy, slang-rich, and code-mixed expressions commonly found in Indonesian social media hate speech.

IndoRoBERTa is a RoBERTa-based encoder pretrained on broader Indonesian text corpora, particularly the Indonesian portion of the OSCAR web corpus. It uses byte-pair encoding with a vocabulary of approximately 50,000 tokens and preserves letter case during tokenization. Compared with IndoBERTweet, IndoRoBERTa is better aligned with more standardized and formal written Indonesian. These differences in pretraining data and tokenization motivate using both

models as complementary base encoders in the ensemble.

Both models are fine-tuned independently using the base-model training subset described in Section 2.1. Text preprocessing is deliberately minimal. Because the Indonesian Hate Speech Superset already replaces usernames and URLs with fixed placeholder tokens, no additional manual normalization, slang substitution, or stop-word removal is applied. Each post is passed directly to the corresponding model's subword tokenizer. IndoBERTweet uses a lowercased WordPiece tokenizer (`do_lower_case = True`), whereas IndoRoBERTa preserves case through its byte-pair encoding tokenizer. Sequences are truncated or padded to a maximum length of 128 tokens, and each model is extended with a sequence classification head configured for binary classification.

Both base Transformer models are fine-tuned using the standard unweighted cross-entropy loss. No class weighting, focal loss, label smoothing, data augmentation, resampling, or class rebalancing is applied during training. This design isolates the contribution of the ensemble strategy from training-level imbalance handling. It ensures that any observed improvement is not attributable to modified loss functions or altered class distributions. The complete training configuration, including optimizer settings, scheduler, warmup steps, gradient clipping, loss function, checkpoint selection criterion, and random seeds, is summarized in Table 2.

Table 2. Training configuration for Transformer-based models

Parameter	Value
Pretrained models	IndoBERTweet, IndoRoBERTa
Tokenization	Model-specific subword tokenizer
Maximum sequence length	128 tokens
Truncation strategy	Right-side truncation
Batch size	16 per device
Gradient accumulation	None
Number of epochs	3
Learning rate	2×10^{-5}
Optimizer	AdamW
AdamW $\beta_1, \beta_2, \epsilon$	0.9, 0.999, 1×10^{-8}
Weight decay	0.01
Learning rate scheduler	Linear decay
Warmup steps	161 steps, computed as 10% of total training steps
Gradient clipping	Max gradient norm = 1.0
Loss function	Standard unweighted cross-entropy
Class weighting	None
Focal loss	Not applied
Label smoothing	None
Data augmentation	None
Resampling	None
Evaluation strategy	Epoch-level evaluation
Checkpoint selection	Best validation macro-F1
Early stopping patience	2 epochs
Data split seed	42
Training seeds	42, 1, 2, 3, 4

After fine-tuning, each base model produces class probability distributions via the softmax output layer for both the meta-training validation set and the held-out test set. These probability outputs are subsequently used

as inputs for the ensemble strategies described in Sections 2.3 and 2.4.

2.3 Ensemble Strategies and Probability Aggregation

This study investigates two ensemble learning strategies that operate at the probability level: a simple soft-voting ensemble and a stacking-based ensemble. Both approaches are designed to combine the outputs of heterogeneous Transformer-based base models without resorting to data-level balancing techniques.

The primary ensemble approach investigated in this study is a soft-voting ensemble that combines the probabilistic outputs of the fine-tuned base models. For a given input instance, IndoBERTweet and IndoRoBERTa each output a probability distribution over the two class labels. The final ensemble prediction is obtained by averaging the predicted class probabilities from the base models and selecting the class with the highest averaged probability, as formally defined in Equation 1.

Given an input text x and K base models, the soft-voting prediction is defined as Equation 1.

$$\hat{y} = \arg \max_{c \in \{0,1\}} \frac{1}{K} \sum_{k=1}^K P_k(c | x) \quad (1)$$

In Equation 1, K denotes the number of base models. $P_k(c | x)$ represents the probability assigned to class c by the k -th model, and $\hat{y}$ is the final predicted label obtained by averaging probabilities. In this study, $K = 2$, corresponding to IndoBERTweet and IndoRoBERTa.

The two base models are combined using static equal weights, with $w_1 = w_2 = 0.5$. No weight optimization is performed. Therefore, Equation 1 is equivalent to a uniform arithmetic mean of the class-probability outputs produced by the two models. This design keeps the soft-voting strategy parameter-free and avoids validation-set weight tuning that could introduce additional model-selection bias or data leakage.

This probability-level integration enables the ensemble to leverage complementary confidence patterns between the base models while remaining computationally efficient. Unlike hard voting, which relies solely on discrete class predictions, soft voting incorporates confidence information and can provide more stable predictions under imbalanced class distributions. Importantly, this approach does not require additional training stages or modification of the original data distribution.

In addition to the soft-voting ensemble, a stacking ensemble is implemented as a comparative ensemble strategy. In the stacking framework, the predicted class probabilities from the base models on the meta-training subset serve as input features for a meta-learner, which is trained to learn an optimal combination of the base models' outputs.

In this study, Logistic Regression is employed as the meta-learner. For each input instance, the probability outputs produced by IndoBERTweet and IndoRoBERTa are concatenated to form a low-dimensional feature vector representing the confidence scores of the base models for each class. For binary classification, the concatenated feature vector consists of the predicted probabilities for both classes from each base model, yielding a 4-dimensional input representation for the meta-learner. This probability-level representation serves as input to the Logistic Regression model, which is trained only on the meta-training validation subset, not on the held-out test set.

After training, the learned Logistic Regression meta-model is applied to the corresponding probability features generated from the held-out test set. This ensures that the stacking ensemble is evaluated on unseen data and prevents information leakage between training and evaluation phases. The stacking ensemble is therefore used as a trained comparison point for soft voting, rather than as the primary proposed method.

Logistic Regression is selected as the meta-learner due to its stability, interpretability, and suitability for low-dimensional feature spaces. In particular, the learned coefficients provide insight into the relative contribution of each base model to the final decision. Within this study, the stacking ensemble serves two purposes: first, as a performance comparison with the simpler soft-voting ensemble; and second, as an analytical tool for examining how the complementary base models influence the ensemble's predictions under imbalanced data conditions.

2.4 Baseline Models and Evaluation Metrics

To provide a comprehensive performance comparison, this study evaluates the proposed ensemble methods against a traditional machine learning baseline and individual Transformer-based base models. All implemented baseline models are evaluated using the same held-out test set to ensure consistency and fairness.

As a traditional baseline, a Support Vector Machine (SVM) classifier with TF-IDF features is implemented. The TF-IDF representation is constructed from the raw text using a fixed vocabulary size, and the SVM is trained using a linear kernel. The SVM baseline serves as a representative non-neural approach commonly used in earlier hate speech detection studies and provides a reference point for assessing the performance gains achieved by Transformer-based models and ensemble strategies.

Model performance is evaluated using multiple classification metrics, with particular emphasis on metrics suitable for imbalanced data. Accuracy is reported to provide an overall measure of correctness. Precision, recall, and F1-score are reported at the macro level, and the macro-averaged F1-score is used as the primary evaluation metric, as it assigns equal weight to each class and provides a more reliable assessment of

model performance under class imbalance. Per-class F1 scores are also reported for the Non-Hate and Hate classes to assess class-specific performance. For Transformer-based models and ensemble methods, results are reported as the mean $\pm$ standard deviation across five training seeds (42, 1, 2, 3, and 4), using a fixed stratified data split.

To assess whether observed performance differences are statistically reliable, paired significance tests are conducted on the held-out test predictions. Specifically, McNemar's test with Yates' correction for continuity is used to compare paired classification outcomes. In contrast, a two-sided paired bootstrap test with 10,000 resamples is used to estimate the significance of differences in macro-F1. Statistical testing is performed on the canonical seed, selected as the run whose soft-voting macro-F1 is closest to the five-seed mean, and cross-seed consistency is reported as a supplementary stability check.

In addition to threshold-based classification metrics, ranking and calibration metrics are reported. ROC-AUC and average precision are computed with bootstrap 95% confidence intervals. Because the proposed soft-voting method averages class probabilities, probability calibration is assessed using the Expected Calibration Error (ECE) and the Brier score. Finally, to complement aggregate quantitative metrics, false-negative and false-positive predictions are examined through a qualitative error taxonomy. All evaluation procedures are performed exclusively on the held-out test set.

3. Results and Discussions

This section presents the experimental results of the proposed ensemble approach on the held-out test set. The analysis covers overall classification performance, paired statistical significance testing, ensemble behavior, probability calibration, and qualitative error patterns, with emphasis on macro-averaged metrics suitable for imbalanced classification.

3.1 Overall Performance Comparison Across Models

The overall performance of all evaluated models is summarized in Tables 3a–3c, covering macro-level metrics, per-class F1 scores, ROC-AUC, and average precision. Consistent with the dataset's moderate class imbalance, the analysis emphasizes macro-F1, macro-recall, per-class F1, and ranking-based metrics rather than accuracy alone.

The TF-IDF SVM baseline performs below the Transformer-based models, confirming the benefit of contextual representations for informal Indonesian social media text. Among the individual Transformer models, IndoBERTweet outperforms IndoRoBERTa, which is consistent with its Twitter-domain pretraining and better alignment with slang-rich social media language.

The two ensemble methods obtain the strongest descriptive results, with SoftVoting achieving the highest mean macro-F1, macro-recall, per-class F1, and average precision, while matching Stacking in ROC-AUC. However, the margins among SoftVoting, IndoBERTweet, and Stacking are small, so these descriptive results should be interpreted together with the significance tests in Section 3.2.

Table 3a. Macro-level classification performance of evaluated models

Model	Accuracy	Prec-macro	Rec-macro	F1-macro
SVM + TF-IDF	0.8445	0.8414	0.8392	0.8402
IndoBERTweet	0.8940 $\pm$.0017	0.8909 $\pm$.0020	0.8927 $\pm$.0017	0.8917 $\pm$.0018
IndoRoBERTa	0.8822 $\pm$.0029	0.8788 $\pm$.0029	0.8806 $\pm$.0032	0.8796 $\pm$.0034
Stacking	0.8981 $\pm$.0016	0.8950 $\pm$.0018	0.8971 $\pm$.0016	0.8959 $\pm$.0018
SoftVoting	0.9003 $\pm$.0029	0.8974 $\pm$.0029	0.8990 $\pm$.0032	0.8981 $\pm$.0033

Table 3b. Per-class F1 scores on the held-out test set

Model	F1 (Non-Hate)	F1 (Hate)
SVM + TF-IDF	0.8664	0.8140
IndoBERTweet	0.9075 $\pm$ 0.0018	0.8758 $\pm$ 0.0019
IndoRoBERTa	0.8972 $\pm$ 0.0025	0.8620 $\pm$ 0.0036
Stacking	0.9110 $\pm$ 0.0015	0.8808 $\pm$ 0.0018
SoftVoting	0.9132 $\pm$ 0.0025	0.8831 $\pm$ 0.0035

Table 3c. ROC-AUC and average precision with bootstrap 95% confidence intervals

Model	ROC-AUC	95% CI	AP (PR)	95% CI
SVM + TF-IDF	0.9149	[0.9044, 0.9250]	0.8860	[0.8679, 0.9029]
IndoBERTweet	0.9577	[0.9507, 0.9643]	0.9447	[0.9345, 0.9540]
IndoRoBERTa	0.9522	[0.9451, 0.9592]	0.9359	[0.9246, 0.9464]
Stacking	0.9632	[0.9572, 0.9689]	0.9499	[0.9401, 0.9589]
SoftVoting	0.9632	[0.9571, 0.9689]	0.9500	[0.9401, 0.9589]

The visual results further support the tabular findings. Figure 2 presents the confusion matrices of the evaluated models in both raw-count and row-normalized forms, allowing the per-class recall

behavior of each model to be inspected directly. This is particularly important for the Hate class, where missed detections have practical implications for moderation systems. Figure 3 presents the precision–recall and

ROC curves with bootstrap 95% confidence intervals, complementing the ROC-AUC and average precision values reported in Table 3c.

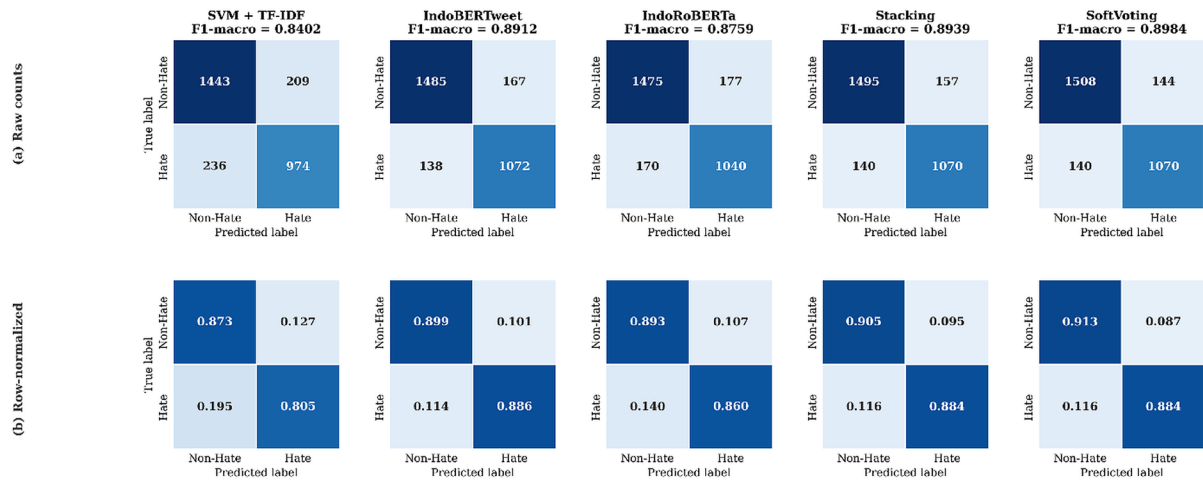


Figure 2. Raw-count and row-normalized confusion matrices of the evaluated models on the held-out test set using the canonical seed. The row-normalized matrices represent per-class recall.

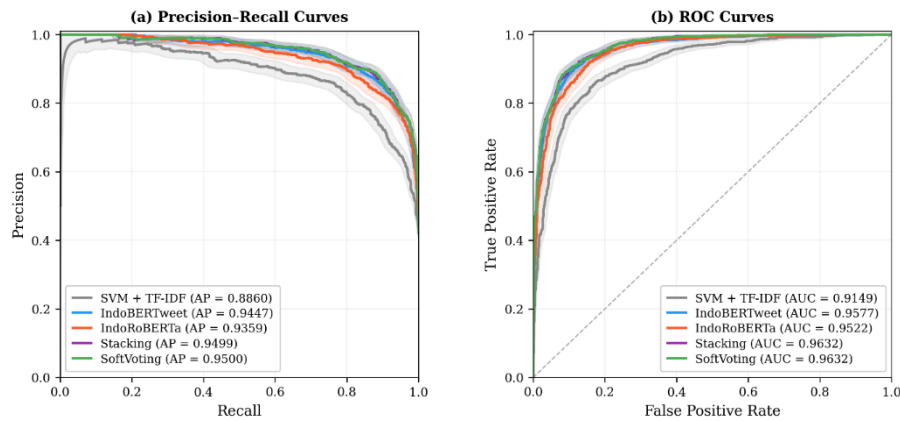


Figure 3. Precision–recall, and ROC curves of the evaluated models on the held-out test set using the canonical seed. Shaded bands indicate bootstrap 95% confidence intervals.

3.2 Statistical Significance Analysis

Paired significance tests were conducted to assess whether the observed performance differences were statistically reliable. McNemar’s test with Yates’ correction was used for paired classification outcomes, while a two-sided paired bootstrap test with 10,000 resamples was used for macro-F1 differences. Testing was performed on the canonical seed, defined as the run whose SoftVoting macro-F1 was closest to the five-seed mean, with cross-seed consistency reported as an additional stability check.

As shown in Table 4, SoftVoting significantly outperforms both the TF-IDF SVM baseline and IndoRoBERTa, with consistent improvements across all five seeds. However, its differences from IndoBERTweet and the trained stacking ensemble are not statistically significant. Therefore, SoftVoting should be interpreted as a simple and competitive ensemble strategy rather than as a method that consistently surpasses the strongest individual or trained ensemble model.

Table 4. Statistical significance results based on paired bootstrap and McNemar tests

Comparison	Δ F1-macro	Bootstrap p-value	McNemar p-value	Significant	Cross-seed consistency
SoftVoting vs SVM + TF-IDF	+0.058	<0.001	<0.001	Yes	5/5
SoftVoting vs IndoRoBERTa	+0.022	<0.001	<0.001	Yes	5/5
SoftVoting vs IndoBERTweet	+0.007	0.064	0.065	No	2/5
SoftVoting vs Stacking	+0.004	0.083	0.093	No	1/5
Stacking vs IndoBERTweet	+0.003	0.352	0.389	No	1/5
IndoBERTweet vs IndoRoBERTa	+0.015	0.010	0.013	Yes	4/5

3.3 Ensemble Behavior Analysis

Beyond aggregate performance metrics, the behavior of the ensemble models can be further examined through the Logistic Regression coefficients learned by the stacking meta-learner. Since the stacking model is trained on the probability outputs of IndoBERTweet and IndoRoBERTa, these coefficients provide an interpretable indication of how the meta-learner weights the contribution of each base encoder. As shown in Figure 4, both base models receive positive coefficients, indicating that the stacking ensemble uses probability information from both pretrained encoders rather than relying on only one model.

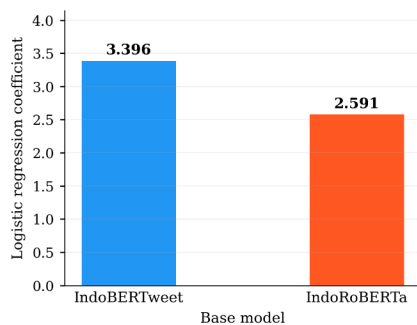


Figure 4. Logistic Regression coefficients of the stacking meta-learner fitted on validation-set probability outputs.

IndoBERTweet receives a larger coefficient (3.396) than IndoRoBERTa (2.591), which is consistent with its stronger individual performance reported in Section 3.1. This suggests that Twitter-domain pretraining of IndoBERTweet provides a stronger signal for detecting hate speech on Indonesian social media. Nevertheless, the positive coefficient assigned to IndoRoBERTa indicates that its probability outputs still contribute useful complementary information. This finding helps explain why equal-weight SoftVoting remains competitive with the trained stacking ensemble: much of the ensemble benefit appears to come from combining complementary probability patterns from

the two encoders, rather than from learning a substantially more complex weighting scheme.

3.4 Probability Calibration Analysis

Because the proposed SoftVoting ensemble averages predicted class probabilities, calibration analysis is used to assess the reliability of its probability estimates. Calibration is evaluated using Expected Calibration Error (ECE) with 15 equal-width bins and the Brier score, as reported in Table 5.

The individual Transformer models show higher calibration errors than the ensemble methods. SoftVoting reduces ECE compared with both IndoBERTweet and IndoRoBERTa, suggesting that equal-weight probability averaging helps moderate the calibration errors of the base models. Stacking achieves the lowest ECE, which is expected because its Logistic Regression meta-learner applies an additional learned transformation to the probability outputs. However, SoftVoting obtains the lowest Brier score, indicating that it remains highly competitive in overall probabilistic accuracy despite not achieving the lowest ECE.

Table 5. Probability calibration results on the held-out test set

Model	ECE ($\downarrow$)	Brier score ($\downarrow$)
Stacking	0.0295	0.0803
SVM + TF-IDF	0.0320	0.1140
SoftVoting	0.0530	0.0801
IndoBERTweet	0.0815	0.0925
IndoRoBERTa	0.0861	0.1001

The reliability diagrams in Figure 5 visually supports these findings. The base Transformer models deviate more from the diagonal calibration line, while SoftVoting shows improved alignment. Stacking provides the closest calibration in terms of ECE, but it requires an additional meta-learning stage. Overall, the calibration results support the interpretation that SoftVoting improves probability reliability over the individual base models while remaining training-free.

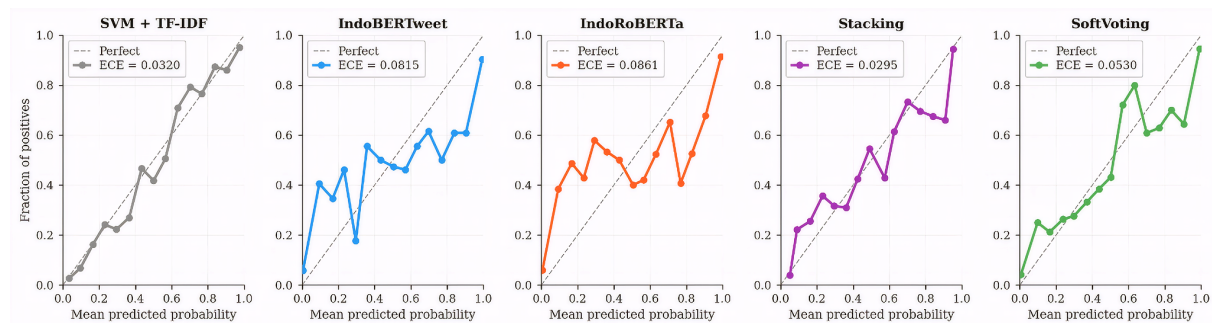


Figure 5. Reliability diagrams of the evaluated models on the held-out test set using 15 equal-width bins.

3.5 Qualitative Error Analysis

To complement the quantitative evaluation, misclassified samples from the SoftVoting ensemble were manually inspected. A random sample of 75 false-negative (FN) and 75 false-positive (FP) predictions was categorized using a non-exclusive error taxonomy, meaning that one instance could belong to more than

one category. The distribution is reported in Table 6 and visualized in Figure 6.

The FN cases are dominated by Political/General framing and Implicit Hate, indicating that the model still struggles when hateful meaning is conveyed indirectly through political commentary, contextual insinuation, or culturally specific references rather than

explicit slurs. In contrast, the FP cases are dominated by Political/General framing and Profanity without protected target, suggesting that the model sometimes

over-interprets adversarial political discourse, general insults, or profanity as hate speech even when no protected group is clearly targeted.

Table 6. Error category distribution in sampled false-negative and false-positive predictions

Error category	FN count	FN (%)	FP count	FP (%)
Explicit Hate	19	25.3	4	5.3
Implicit Hate	31	41.3	15	20.0
Sarcasm/Irony	1	1.3	6	8.0
Code-Mixing	3	4.0	3	4.0
Profanity without protected target	14	18.7	23	30.7
Political/General framing	40	53.3	51	68.0
Target Ambiguity	7	9.3	7	9.3
Possible Label Noise	25	33.3	33	44.0
Other	0	0.0	0	0.0

Note. Categories are non-exclusive; therefore, percentages may sum to more than 100% within each error type. Percentages are computed from 75 sampled false-negative and 75 sampled false-positive predictions.

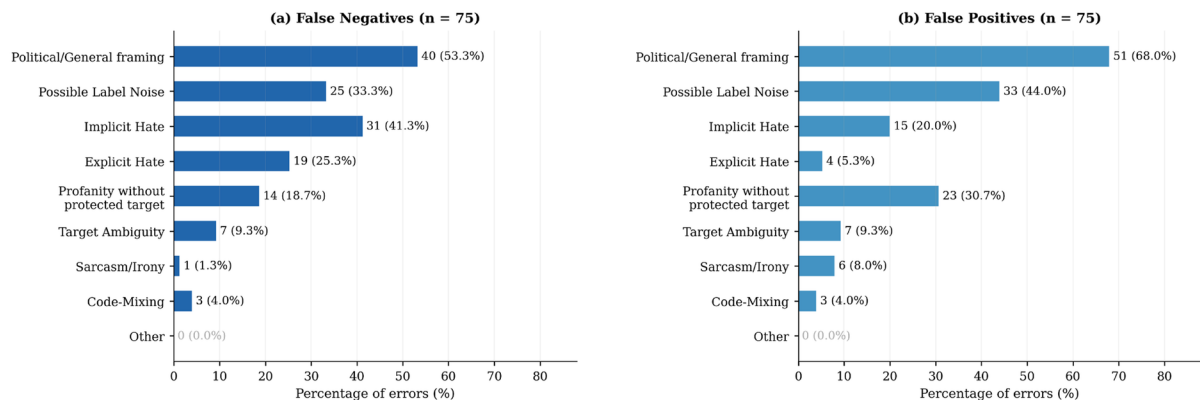


Figure 6. Distribution of qualitative error categories in sampled false-negative and false-positive predictions.

Possible Label Noise also appears frequently in both FN and FP samples, reflecting the difficulty of using a consolidated multi-source benchmark with different annotation schemes. Some apparent errors may therefore reflect ambiguity or inconsistency in the gold labels rather than purely incorrect model behavior.

Code-mixing appears less frequently, but it remains analytically relevant because it can reveal differences between the social-media-specific sensitivity of IndoBERTweet and the broader linguistic coverage of IndoRoBERTa. Representative examples of these dominant error patterns are presented in Table 7.

Table 7. Representative qualitative error examples

No.	Error type	Example excerpt	Model behavior	Interpretation
1	FN: Implicit hate in political framing	“Mereka berbicara era Habibie hingga SBY ... sejak TKA RRC masuk secara masif ...”	True label: hate; predicted: non-hate; SoftVoting p(hate) = 0.005	The text constructs an implicit threat narrative against Chinese foreign workers, but the hostility is embedded in political commentary rather than expressed through an explicit slur.
2	FN: Short ethnic slur	“Males gua bacanya onta”	True label: hate; predicted: non-hate; SoftVoting p(hate) = 0.008	The word “onta” can function as a derogatory ethnic or religious slur. However, the utterance is very short and lacks sufficient context for the ensemble to classify it as hate speech.
3	FN: Register-specific identity slang	“Tae. Isinya banci semwa yak?”	True label: hate; predicted: non-hate; SoftVoting p(hate) = 0.266	IndoBERTweet assigns a higher hate probability than IndoRoBERTa, suggesting that the Twitter-pretrained model is more sensitive to informal identity-based slang. However, equal-weight averaging still keeps the final probability below the decision threshold.
4	FP: Political idiom misread as literal hate	“Prabowo makin membabi buta. Serang siapa saja.”	True label: non-hate; predicted: hate; SoftVoting p(hate) = 0.992	The idiom “membabi buta” means acting recklessly or indiscriminately, but the model appears to associate the phrase with animal-based insults and adversarial political tone.
5	FP: Profanity as venting without protected target	“Bangad kebaca bangsad”	True label: non-hate; predicted: hate; SoftVoting p(hate) = 0.998	The text contains strong profanity, but no protected identity group is targeted. The model confuses general abusive language or venting with hate speech.
6	FP: Possible label noise	“... para bajingan minoritas cina negeri ini terhadap mayoritas islam ...”	True label: non-hate; predicted: hate; SoftVoting p(hate) = 0.995	The text contains identity-based derogatory language targeting the Chinese-Indonesian minority. The model prediction is plausible, and the non-hate gold label may reflect annotation inconsistency inherited from the source corpus.

The examples in Table 7 show that the remaining errors are driven less by simple lexical mistakes than by more complex linguistic and annotation-related phenomena. False negatives often involve implicit political framing, short or culturally specific slurs, and register-specific slang, whereas false positives commonly arise from political idioms, general profanity, or aggressive but non-identity-based discourse. These findings reinforce the quantitative results: SoftVoting is competitive, but Indonesian hate speech detection remains challenging when harmful meaning is indirect, context-dependent, or inconsistently annotated.

4. Conclusions

This study examined whether a simple, training-free, probability-level ensemble can provide competitive performance for Indonesian hate speech detection under moderate class imbalance without relying on data resampling. By combining IndoBERTweet and IndoRoBERTa through equal-weight SoftVoting, the proposed approach achieved the highest descriptive macro-level performance, with a macro-F1 of 0.8981 ± 0.0033 and a macro-recall of 0.8990 ± 0.0032 . However, the statistical results support a cautious interpretation: SoftVoting significantly outperformed the TF-IDF SVM baseline and IndoRoBERTa, but its differences from IndoBERTweet and the trained stacking ensemble were not statistically significant. Thus, the main contribution of this study lies in demonstrating that equal-weight SoftVoting is a simple, reproducible, and competitive alternative without class re-weighting, loss-function modification, data resampling, or an additional meta-learner.

The calibration analysis showed that SoftVoting improves probability reliability over both individual Transformer models, although Stacking achieved the lowest ECE through an additional learned transformation. The qualitative error analysis further showed that remaining errors are dominated by implicit political framing, culturally specific slang, profanity without a protected target, target ambiguity, and possible label noise. These findings indicate that Indonesian hate speech detection remains challenging when harmful meaning is indirect, context-dependent, or inconsistently annotated.

This study is limited to a single consolidated benchmark with moderate rather than severe class imbalance, fixed equal ensemble weights, native probability calibration, and no direct comparison with algorithm-level imbalance-aware losses such as class-weighted cross-entropy, focal loss, or class-balanced loss. Future research should address these limitations by evaluating the approach across different corpora and social media platforms, comparing model-level ensembling with imbalance-aware training objectives, examining post-hoc calibration methods, and extending the task to multi-class or multi-label hate speech categorization with richer contextual information.

Acknowledgements

This research was financially supported by the Lambung Mangkurat University Research Grant, year 2026, with grant number 079/105/UN8.2/PG/2026.

References

- [1] S. Stieglitz and B. Ross, "The Impact of Social Media on Social Cohesion: A Double-Edged Sword," *MaC*, vol. 10, no. 2, pp. 104–107, May 2022, doi: [10.17645/mac.v10i2.5792](https://doi.org/10.17645/mac.v10i2.5792).
- [2] J. B. Walther, "Social media and online hate," *Current Opinion in Psychology*, vol. 45, p. 101298, Jun. 2022, doi: [10.1016/j.copsyc.2021.12.010](https://doi.org/10.1016/j.copsyc.2021.12.010).
- [3] M. Retta, "A pragmatic and discourse analysis of hate words on social media," *Internet Pragmatics*, vol. 6, no. 2, pp. 197–218, Dec. 2023, doi: [10.1075/ip.00096.ret](https://doi.org/10.1075/ip.00096.ret).
- [4] D. Hartmann, A. Oueslati, D. Staufer, L. Pohlmann, S. Munzert, and H. Heuer, "Lost in Moderation: How Commercial Content Moderation APIs Over- and Under-Moderate Group-Targeted Hate Speech and Linguistic Variations," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, Yokohama Japan: ACM, Apr. 2025, pp. 1–26. doi: [10.1145/3706598.3713998](https://doi.org/10.1145/3706598.3713998).
- [5] J. Gracia-Calandín and L. Suárez-Montoya, "The eradication of hate speech on social media: a systematic review," *JICES*, vol. 21, no. 4, pp. 406–421, Nov. 2023, doi: [10.1108/JICES-11-2022-0098](https://doi.org/10.1108/JICES-11-2022-0098).
- [6] C. Silva, "Fighting Against Hate Speech: A Case for Harnessing Interactive Digital Counter-Narratives," in *Interactive Storytelling*, vol. 14383, L. Holloway-Attaway and J. T. Murray, Eds., in *Lecture Notes in Computer Science*, vol. 14383, Cham: Springer Nature Switzerland, 2023, pp. 159–174. doi: [10.1007/978-3-031-47655-6_10](https://doi.org/10.1007/978-3-031-47655-6_10).
- [7] A. M. Kumar and S. Gupta, "Governance of Social Media Platforms: A Literature Review," *PAJAIS*, vol. 15, pp. 56–86, Mar. 2023, doi: [10.17705/1pais.15103](https://doi.org/10.17705/1pais.15103).
- [8] A. Rawat, S. Kumar, and S. S. Samant, "Hate speech detection in social media: Techniques, recent trends, and future challenges," *WIREs Computational Stats*, vol. 16, no. 2, p. e1648, Mar. 2024, doi: [10.1002/wics.1648](https://doi.org/10.1002/wics.1648).
- [9] F. Alkomah and X. Ma, "A Literature Review of Textual Hate Speech Detection Methods and Datasets," *Information*, vol. 13, no. 6, p. 273, May 2022, doi: [10.3390/info13060273](https://doi.org/10.3390/info13060273).
- [10] A. Gandhi et al., "Hate speech detection: A comprehensive review of recent works," *Expert Systems*, vol. 41, no. 8, p. e13562, Aug. 2024, doi: [10.1111/exsy.13562](https://doi.org/10.1111/exsy.13562).
- [11] E. W. Pamungkas, D. G. P. Putri, and A. Fatmawati, "Hate Speech Detection in Bahasa Indonesia: Challenges and Opportunities," *IJACSA*, vol. 14, no. 6, 2023, doi: [10.14569/IJACSA.2023.01406125](https://doi.org/10.14569/IJACSA.2023.01406125).
- [12] M. O. Ibrohim and I. Budi, "Hate speech and abusive language detection in Indonesian social media: Progress and challenges," *Heliyon*, vol. 9, no. 8, p. e18647, Aug. 2023, doi: [10.1016/j.heliyon.2023.e18647](https://doi.org/10.1016/j.heliyon.2023.e18647).
- [13] H. Margono, M. Saud, and A. Ashfaq, "Dynamics of hate speech in social media: insights from Indonesia," *GKMC*, Aug. 2024, doi: [10.1108/GKMC-11-2023-0464](https://doi.org/10.1108/GKMC-11-2023-0464).
- [14] A. D. Sanya and L. H. Suadaa, "Handling Imbalanced Dataset on Hate Speech Detection in Indonesian Online News Comments," in *2022 10th International Conference on Information and Communication Technology (ICOICT)*, Bandung, Indonesia: IEEE, Aug. 2022, pp. 380–385. doi: [10.1109/ICOICT55009.2022.9914883](https://doi.org/10.1109/ICOICT55009.2022.9914883).
- [15] S. Riyadi, A. Divayu Andriyani, and S. Noraini Sulaiman, "Improving Hate Speech Detection Using Double-Layers Hybrid CNN-RNN Model on Imbalanced Dataset," *IEEE Access*, vol. 12, pp. 159660–159668, 2024, doi: [10.1109/ACCESS.2024.3487433](https://doi.org/10.1109/ACCESS.2024.3487433).
- [16] R. Narula and P. Chaudhary, "A comprehensive review on detection of hate speech for multi-lingual data," *Soc. Netw. Anal. Min.*, vol. 14, no. 1, p. 244, Jan. 2025, doi: [10.1007/s13278-024-01401-y](https://doi.org/10.1007/s13278-024-01401-y).
- [17] J.-G. Gaudreault and P. Branco, "Empirical analysis of performance assessment for imbalanced classification," *Mach*

- Learn, vol. 113, no. 8, pp. 5533–5575, Aug. 2024, doi: [10.1007/s10994-023-06497-5](https://doi.org/10.1007/s10994-023-06497-5).
- [18] C. Padurariu and M. E. Breaban, “Dealing with Data Imbalance in Text Classification,” *Procedia Computer Science*, vol. 159, pp. 736–745, 2019, doi: [10.1016/j.procs.2019.09.229](https://doi.org/10.1016/j.procs.2019.09.229).
- [19] D. Dablain, B. Krawczyk, and N. V. Chawla, “DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data,” *IEEE Trans. Neural Netw. Learning Syst.*, vol. 34, no. 9, pp. 6390–6404, Sep. 2023, doi: [10.1109/TNNLS.2021.3136503](https://doi.org/10.1109/TNNLS.2021.3136503).
- [20] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, “Class-Balanced Loss Based on Effective Number of Samples,” in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA: IEEE, Jun. 2019, pp. 9260–9269, doi: [10.1109/CVPR.2019.00949](https://doi.org/10.1109/CVPR.2019.00949).
- [21] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, “Focal Loss for Dense Object Detection,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice: IEEE, Oct. 2017, pp. 2999–3007, doi: [10.1109/ICCV.2017.324](https://doi.org/10.1109/ICCV.2017.324).
- [22] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On Calibration of Modern Neural Networks,” in *Proceedings of the 34th International Conference on Machine Learning*, D. Precup and Y. W. Teh, Eds., in Proceedings of Machine Learning Research, vol. 70. PMLR, Aug. 2017, pp. 1321–1330. [Online]. Available: <https://proceedings.mlr.press/v70/guo17a.html>
- [23] M. Subramanian, V. Easwaramoorthy Sathiskumar, G. Deepalakshmi, J. Cho, and G. Manikandan, “A survey on hate speech detection and sentiment analysis using machine learning and deep learning models,” *Alexandria Engineering Journal*, vol. 80, pp. 110–121, Oct. 2023, doi: [10.1016/j.aej.2023.08.038](https://doi.org/10.1016/j.aej.2023.08.038).
- [24] S. S. Roy, A. Roy, P. Samui, M. Gandomi, and A. H. Gandomi, “Hateful Sentiment Detection in Real-Time Tweets: An LSTM-Based Comparative Approach,” *IEEE Trans. Comput. Soc. Syst.*, vol. 11, no. 4, pp. 5028–5037, Aug. 2024, doi: [10.1109/TCS.2023.3260217](https://doi.org/10.1109/TCS.2023.3260217).
- [25] S. Nouas, F. Boumahdi, A. Madani, and F. Berrhail, “Cyberbullying: A BERT Bi-LSTM Solution for Hate Speech Detection,” in *Proceedings of the International Conference on Applied Cybersecurity (ACS) 2023*, H. Zantout and H. Ragab Hassen, Eds., in Lecture Notes in Networks and Systems, vol. 760. Cham: Springer Nature Switzerland, 2023, pp. 57–65, doi: [10.1007/978-3-031-40598-3_7](https://doi.org/10.1007/978-3-031-40598-3_7).
- [26] M. Tonneau, D. Liu, S. Fraiberger, R. Schroeder, S. Hale, and P. Röttger, “From Languages to Geographies: Towards Evaluating Cultural Bias in Hate Speech Datasets,” in *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, Mexico City, Mexico: Association for Computational Linguistics, 2024, pp. 283–311, doi: [10.18653/v1/2024.woah-1.23](https://doi.org/10.18653/v1/2024.woah-1.23).
- [27] N. I. Pratiwi, I. Budi, and I. Alfina, “Hate Speech Detection on Indonesian Instagram Comments using FastText Approach,” in *2018 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, Yogyakarta: IEEE, Oct. 2018, pp. 447–450, doi: [10.1109/ICACSIS.2018.8618182](https://doi.org/10.1109/ICACSIS.2018.8618182).
- [28] M. O. Ibrohim and I. Budi, “Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter,” in *Proceedings of the Third Workshop on Abusive Language Online*, Florence, Italy: Association for Computational Linguistics, 2019, pp. 46–57, doi: [10.18653/v1/W19-3506](https://doi.org/10.18653/v1/W19-3506).
- [29] I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata, “Hate speech detection in the Indonesian language: A dataset and preliminary study,” in *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, Bali: IEEE, Oct. 2017, pp. 233–238, doi: [10.1109/ICACSIS.2017.8355039](https://doi.org/10.1109/ICACSIS.2017.8355039).
- [30] A. Muhamad Faza, Y. Sibaroni, and S. S. Prasetyowati, “A Comparative Study on Handling Imbalanced Data in Indonesian Hate Speech Detection Using FastText and BiLSTM,” *ijoint*, vol. 11, no. 2, pp. 136–149, Dec. 2025, doi: [10.21108/ijoint.v11i2.9513](https://doi.org/10.21108/ijoint.v11i2.9513).
- [31] G. Ramos et al., “A comprehensive review on automatic hate speech detection in the age of the transformer,” *Soc. Netw. Anal. Min.*, vol. 14, no. 1, p. 204, Oct. 2024, doi: [10.1007/s13278-024-01361-3](https://doi.org/10.1007/s13278-024-01361-3).
- [32] A. Albladi et al., “Hate Speech Detection Using Large Language Models: A Comprehensive Review,” *IEEE Access*, vol. 13, pp. 20871–20892, 2025, doi: [10.1109/ACCESS.2025.3532397](https://doi.org/10.1109/ACCESS.2025.3532397).
- [33] P. A. Mufva, K. H. Chandra, K. F. Aji, I. A. Iswanto, and S. Joddy, “Performance comparison of deep learning approaches for Indonesian twitter hate speech detection using IndoBERTweet embedding,” *Procedia Computer Science*, vol. 269, pp. 1663–1671, 2025, doi: [10.1016/j.procs.2025.09.109](https://doi.org/10.1016/j.procs.2025.09.109).
- [34] M. Ridha, D. Nurjanah, and M. Rakha, “Multilabel Classification Abusive Language and Hate Speech on Indonesian Twitter Using Transformer Model: IndoBERTweet & IndoRoBERTa,” in *2024 International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA)*, Bali, Indonesia: IEEE, Dec. 2024, pp. 48–54, doi: [10.1109/ICICyTA64807.2024.10912874](https://doi.org/10.1109/ICICyTA64807.2024.10912874).
- [35] F. Koto, J. H. Lau, and T. Baldwin, “IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 10660–10668, doi: [10.18653/v1/2021.emnlp-main.833](https://doi.org/10.18653/v1/2021.emnlp-main.833).
- [36] “flax-community/indonesian-roberta-base · Hugging Face.” Accessed: Mar. 30, 2026. [Online]. Available: <https://huggingface.co/flax-community/indonesian-roberta-base>
- [37] A. M. U. D. Khanday, S. T. Rabani, Q. R. Khan, and S. H. Malik, “Detecting twitter hate speech in COVID-19 era using machine learning and ensemble learning techniques,” *International Journal of Information Management Data Insights*, vol. 2, no. 2, p. 100120, Nov. 2022, doi: [10.1016/j.jjime.2022.100120](https://doi.org/10.1016/j.jjime.2022.100120).
- [38] P. Garg, M. K. Sharma, and P. Kumar, “Improving Hate Speech Classification Through Ensemble Learning and Explainable AI Techniques,” *Arab J Sci Eng*, vol. 50, no. 15, pp. 11631–11644, Aug. 2025, doi: [10.1007/s13369-024-09540-2](https://doi.org/10.1007/s13369-024-09540-2).
- [39] O. Sagi and L. Rokach, “Ensemble learning: A survey,” *WIREs Data Min. & Knowl.*, vol. 8, no. 4, p. e1249, Jul. 2018, doi: [10.1002/widm.1249](https://doi.org/10.1002/widm.1249).
- [40] T. Huang, Q. She, and J. Zhang, “BoostingBERT: Integrating Multi-Class Boosting into BERT for NLP Tasks,” 2020, *arXiv*, doi: [10.48550/ARXIV.2009.05959](https://doi.org/10.48550/ARXIV.2009.05959).
- [41] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan, “Ensemble deep learning: A review,” *Engineering Applications of Artificial Intelligence*, vol. 115, p. 105151, Oct. 2022, doi: [10.1016/j.engappai.2022.105151](https://doi.org/10.1016/j.engappai.2022.105151).
- [42] A. Mohammed and R. Kora, “An effective ensemble deep learning framework for text classification,” *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 10, pp. 8825–8837, Nov. 2022, doi: [10.1016/j.jksuci.2021.11.001](https://doi.org/10.1016/j.jksuci.2021.11.001).
- [43] M. U. Salur and İ. Aydın, “A soft voting ensemble learning-based approach for multimodal sentiment analysis,” *Neural Comput. & Applic.*, vol. 34, no. 21, pp. 18391–18406, Nov. 2022, doi: [10.1007/s00521-022-07451-7](https://doi.org/10.1007/s00521-022-07451-7).
- [44] P. Narwariya and S. K. Tiwari, “A robust multi-class text classification framework using TF-IDF features and soft voting ensembles,” in *Recent Advances in Computational Methods in Science and Technology*, CRC Press, 2026.