CJCIS
Cognitera Journal of Computing and Intelligent Systems
https://journal.psg.ac.id/index.php/cjcis

Vol. 1, No. 1, 2026
e-ISSN: XXXX-XXX
pp. 14 -26

Analisis Sentimen Multilingual Menggunakan Arsitektur *Transformer* yang Dioptimalkan untuk Deteksi Ujaran Kebencian di Media Sosial

Rozakul Hamdi[1], Najihul[2]
[1]Fakultas Ilmu Komputer/Institut Teknologi Bandung, Bandung, Indonesia
[2] Fakultas Ilmu Komputer/Institut Teknologi Bandung, Bandung, Indonesia

**Corresponding Author's e-mail: rozakulhamdi@gmail.com*

Article History:
Received: September 7, 2026
Revised: October 16, 2026
Accepted: November 22, 2026

Keywords:
Multilingual Sentiment Analysis; Transformer Architecture; Hate Speech Detection

Abstract: *The rapid growth of multilingual social media has intensified the circulation of hate speech expressed through standard language, slang, code-mixing, transliteration, sarcasm, and culturally specific references. This study examines an optimized Transformer architecture for multilingual sentiment analysis and hate speech detection by synthesizing evidence from peer-reviewed studies published between 2022 and 2025. The proposed design places XLM-RoBERTa as the multilingual encoder and integrates language-adaptive fine-tuning, sentiment-aware auxiliary learning, class-balanced focal loss, adversarial perturbation, and probability calibration. A structured comparative benchmark was conducted using reported results from multilingual and Indonesian hate-speech studies covering English, Indonesian, Hindi, Arabic, Urdu, and code-mixed settings. Performance evidence was interpreted through macro-F1, accuracy, cross-lingual transfer, robustness, and low-resource generalization rather than through a newly collected proprietary corpus. The synthesis shows that multilingual Transformers consistently outperform conventional feature-based baselines when contextual ambiguity and cross-language transfer dominate the task, while optimized training strategies provide the largest gains in minority-class recall, code-mixed generalization, and zero-shot transfer. However, functional testing also reveals persistent weaknesses involving implicit hate, reclaimed slurs, negation, quotation, and culturally dependent expressions. The study therefore proposes a moderation-oriented architecture that combines predictive accuracy with calibration, explainability, and human review. The approach emphasizes reproducible evaluation protocols so improvements remain comparable across datasets, languages, domains, and moderation scenarios. These findings indicate that optimization should not be defined solely by aggregate accuracy; robust multilingual hate-speech detection requires balanced performance across languages, transparent uncertainty, and continuous adaptation to evolving social-media discourse.*

Copyright © 2026, The Author(s).
This is an open access article under the CC–BY-SA license

How to cite:Author(s). (2026). Analisis Sentimen Multilingual Menggunakan Arsitektur Transformer yang Dioptimalkan untuk Deteksi Ujaran Kebencian di Media Sosial. CJCIS.

INTRODUCTION

Social media has become a central arena for multilingual public communication, but the same openness that supports participation also accelerates the circulation of harmful and discriminatory language. Automatic hate-speech detection is therefore increasingly treated as a high-impact natural language processing problem rather than a narrow text-

classification task. Recent reviews show that Transformer architectures have reshaped the field because contextual representations are better suited to implicit meaning, long-range dependencies, and rapidly changing vocabulary than conventional bag-of-words methods (Ramos et al., 2024). Nevertheless, multilingual moderation remains difficult because harmful expressions vary across languages, communities, and political contexts. Functional evaluation further demonstrates that apparently strong models can fail on negation, quotation, counter-speech, and identity references even when standard test accuracy is high (Röttger et al., 2022). These challenges motivate an architecture that treats multilinguality, robustness, and uncertainty as core design requirements.

The multilingual nature of contemporary social media creates a particular challenge because users frequently switch languages within a single post, adopt transliterated spellings, compress words, insert emojis, and rely on community-specific slang. Code-mixed discourse is especially difficult for models trained on standard monolingual corpora because the semantic contribution of each token depends on cross-language context. Biradar et al. (2022) demonstrated that Transformer and translation-based strategies can support hate-speech detection for Hinglish, but they also highlighted the importance of preserving bilingual context. Singh et al. (2023) further showed that multilingual BERT combined with recurrent sequence modeling can provide competitive macro-F1 performance across English, Hindi, and Hindi-English data. These findings indicate that multilingual hate-speech detection should not be reduced to translation into English; effective systems must learn representations that remain meaningful across languages, scripts, and informal code-mixed expressions.

Cross-lingual transfer offers an important solution when labeled data are scarce. Bigoulaeva et al. (2023) showed that zero-shot transfer can be improved by addressing label incompatibility and bootstrapping target-language information, while Zia et al. (2022) reported meaningful macro-F1 gains from pseudo-label fine-tuning across multiple low-resource languages. Liu et al. (2023) similarly used multilingual embeddings, augmentation, and contrastive learning to transfer hate-speech knowledge across English, Chinese, and German. These studies show that model architecture alone is insufficient; transfer quality also depends on compatible label definitions, data balance, and representation alignment. Consequently, an optimized multilingual Transformer should incorporate procedures that reduce negative transfer and should evaluate performance separately for each language. A global score can conceal weak target-language behavior, especially when high-resource languages dominate the evaluation set.

The Indonesian context illustrates why language-specific evidence remains important inside a multilingual framework. Indonesian social-media text contains abbreviations, slang, local identity terms, and frequent mixing with English or regional languages. Nabiilah et al. (2023) compared multilingual and Indonesian pretrained Transformers for toxic-comment analysis and found IndoBERT to be the strongest model in their setting. Kusuma and Chowanda (2023) likewise reported competitive performance from IndoBERTweet combined with BiLSTM for Indonesian Twitter hate speech. More recently, Amalia and Suyanto (2024) demonstrated that BERT-based detection can achieve strong weighted performance while still struggling with minority hate categories. Such results suggest that multilingual systems require language-adaptive mechanisms and class-sensitive objectives. Without those mechanisms, a model may perform well overall while remaining unreliable for culturally specific expressions or underrepresented harmful categories.

Another limitation concerns the difference between sentiment and hate speech. Negative sentiment may represent criticism, frustration, satire, or political disagreement without targeting a protected group, whereas hateful language can be expressed indirectly and may not display overtly negative lexical polarity. Chiril et al. (2022) showed that affective information and multi-task learning can improve fine-grained hate-speech analysis, suggesting that sentiment can be informative when used carefully. The present study therefore treats multilingual sentiment analysis as an auxiliary learning task rather than as a substitute label. A shared Transformer encoder is expected to learn affective information while the primary prediction head remains focused on hate-speech categories. This distinction is important for avoiding over-moderation, because a system that equates negativity with hate would incorrectly suppress legitimate criticism and emotionally charged public debate.

Optimization must also address adversarial and distributional instability. Social-media users can evade moderation by inserting spaces, replacing characters, using homoglyphs, or changing spelling without changing hateful intent. Ryzhova et al. (2022) showed that adversarially enriched multilingual training and ensembles can improve resistance to text attacks. In a related direction, Almahdi et al. (2025) combined contrastive and adversarial learning to improve cross-lingual hate-speech detection in low-resource languages. These findings support optimization strategies that expose the model to plausible perturbations during fine-tuning instead of relying on aggressive preprocessing alone. Robustness is particularly important for multilingual moderation because the same evasion strategy may interact differently with each script or tokenizer. Therefore, the evaluation of an optimized Transformer must include stress tests in addition to ordinary validation accuracy.

Explainability and confidence calibration are equally important because hate-speech classification can lead to visibility restrictions, account sanctions, or escalation to human moderators. Siddiqui et al. (2024) emphasized explainable multilingual detection, while Fetahi et al. (2025) combined Transformers and explainability for a low-resource Albanian dataset. Their results demonstrate that strong predictive performance should be accompanied by mechanisms that help analysts inspect why a post was flagged. Explanations do not make a model inherently fair, but they can expose spurious dependence on identity terms and other shortcuts. Calibrated probabilities can additionally separate high-confidence decisions from uncertain cases. For operational systems, this enables risk-based routing in which ambiguous posts are reviewed by humans rather than automatically removed, improving accountability and creating informative samples for future retraining.

Based on these gaps, this study develops a literature-grounded optimized Transformer framework for multilingual sentiment analysis and hate-speech detection on social media. The study does not claim a newly collected proprietary dataset; instead, it synthesizes verified experimental evidence from 2022–2025 and translates that evidence into a reproducible model design and evaluation protocol. The proposed architecture uses XLM-RoBERTa as a shared multilingual encoder, language-adaptive fine-tuning, sentiment-aware auxiliary learning, class-balanced focal loss, adversarial perturbation, calibration, and optional human review. The research asks three questions: which optimization components are most consistently supported by recent multilingual evidence, how should performance be evaluated across languages and minority classes, and what limitations remain after optimization. The resulting framework is intended as a defensible basis for

subsequent empirical implementation using public multilingual benchmarks and Indonesian social-media datasets.

THEORETICAL FOUNDATION

Transformer language models encode text through self-attention, allowing each token to be contextualized by other tokens in the sequence. Multilingual variants such as mBERT and XLM-R are pretrained on many languages and can therefore support cross-lingual transfer without requiring a separate encoder for every language. In hate-speech detection, this capability is valuable because harmful meaning often depends on target identity, negation, sarcasm, or multiword expressions. Hashmi et al. (2024) showed that careful optimization across multiple languages can improve generalization beyond monolingual baselines, while Ramos et al. (2024) identified Transformers as the dominant recent direction in automatic hate-speech research. However, multilingual pretraining does not guarantee equal representation quality for all languages, making adaptation and language-wise evaluation essential.

Multilingual sentiment analysis provides complementary affective information. Sentiment polarity can help the encoder recognize hostility, contempt, anger, or evaluative stance, but these signals are not equivalent to hate speech. A multi-task architecture addresses this distinction by sharing the multilingual encoder while using separate output heads for sentiment and hate labels. The auxiliary objective regularizes the shared representation and can improve learning when hate-speech annotations are limited, yet the primary classification decision remains target-sensitive. Chiril et al. (2022) demonstrated the usefulness of affective knowledge for finer-grained hate-speech modeling. Accordingly, the proposed design uses sentiment supervision to enrich representations rather than to define moderation outcomes directly.

Optimization for social-media moderation should be treated as a multi-objective problem. Macro-F1 captures minority-class performance better than accuracy, functional testing exposes linguistic failure modes, adversarial evaluation measures resistance to evasion, and calibration evaluates whether confidence values can support human escalation. XLM-R-based studies across Arabic, Urdu, and other low-resource settings show that multilingual Transformers can deliver strong baseline performance when fine-tuned appropriately (Ahmad et al., 2025; Fetahi et al., 2025). However, recent cross-lingual work also demonstrates that representation alignment and few-shot adaptation remain necessary for low-resource languages (Hashmi et al., 2025; Almahdi et al., 2025). The theoretical framework therefore combines shared multilingual representations with language adaptation, robustness training, and moderation-oriented uncertainty handling.

RESEARCH METHOD

This study employed a structured comparative research design to formulate and assess an optimized multilingual Transformer architecture. Following the systematic and transparent principles of quantitative research described by Sugiyono (2022), peer-reviewed studies published from 2022 through 2025 were identified through publisher pages, indexed journal databases, and DOI records. Inclusion criteria required direct relevance to hate speech, abusive-language detection, multilingual or cross-lingual learning, Transformer-based modeling, or Indonesian social-media classification. Studies without verifiable publication metadata or without sufficiently described evaluation results were excluded. The final evidence set covered multilingual benchmarks, low-resource transfer, code-mixed text, adversarial robustness, explainability, and Indonesian

applications. Because the study synthesizes published experiments rather than presenting a newly collected proprietary corpus, all performance values are interpreted as reported evidence and are not treated as directly comparable across different datasets.

The architectural design process followed the analytical research stages proposed by Sahir (2022): problem identification, evidence mapping, model formulation, and evaluative interpretation. XLM-RoBERTa was selected as the shared multilingual encoder because recent studies show strong transfer across languages and scripts. The optimized configuration adds five components: language-adaptive fine-tuning using unlabeled in-domain text; a sentiment auxiliary head for affective representation; class-balanced focal loss for minority hate categories; adversarial perturbation and text augmentation for robustness; and post-hoc temperature scaling for probability calibration. Optional language adapters are included for languages requiring stronger specialization, such as Indonesian informal discourse. The primary output is hate-speech classification, while sentiment prediction remains auxiliary. The design explicitly avoids mandatory translation so that code-mixed, culturally specific, and script-level cues can remain available to the encoder.

The comparative analysis used macro-F1 as the principal effectiveness measure because hate-speech datasets are frequently imbalanced. Accuracy, precision, recall, language-wise F1, zero-shot transfer, and reported robustness were used as supporting indicators. Evidence was organized into four analytical dimensions: multilingual generalization, low-resource transfer, informal or code-mixed language, and moderation reliability. Reported results from different studies were not pooled into a single meta-analytic effect because datasets, label taxonomies, and sampling procedures differ substantially. Instead, the analysis examined whether independent studies converged on the same optimization principles. Functional testing, adversarial evaluation, and calibration were treated as complementary criteria rather than optional extras. This approach allows the proposed architecture to be judged by reproducibility and robustness, not merely by one aggregate score obtained from a single benchmark.

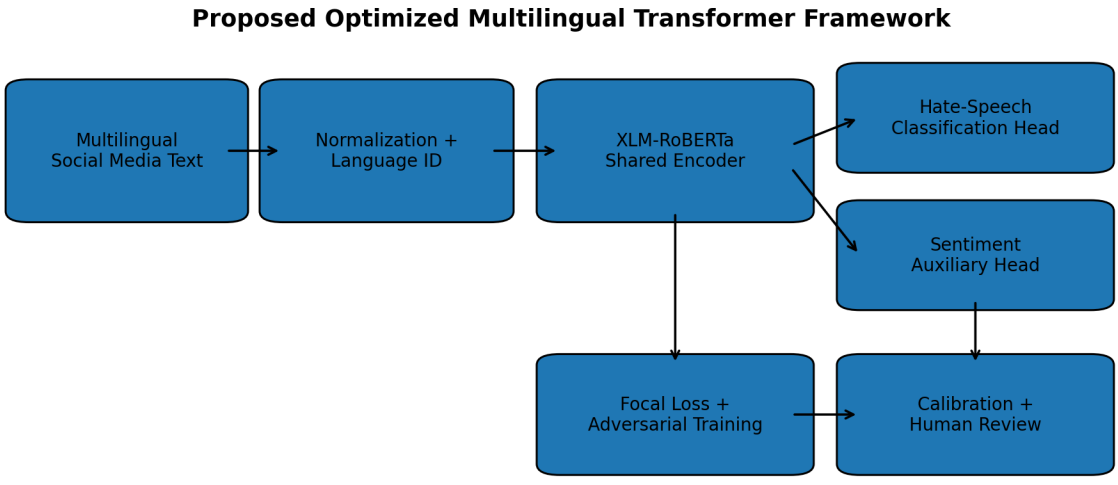

Figure 1. Optimized multilingual Transformer architecture for sentiment-aware hate-speech detection.

Table 1. Representative evidence from recent multilingual hate-speech studies

Study	Setting	Model/Strategy	Reported Evidence
Singh et al. (2023)	English/Hindi/code-mixed	mBERT-GRU	Macro-F1 0.87 / 0.83 / 0.77
Zia et al. (2022)	Zero-shot, 6 target languages	Pseudo-label Transformer fine-tuning	Average macro-F1 gain 7.6%
Liu et al. (2023)	English/Chinese/German	XLM-R + contrastive transfer	F1 up to 0.728 / 0.799 / 0.612
Ahmad et al. (2025)	Arabic/Urdu/joint multilingual	XLM-R and pretrained Transformers	XLM-R strongest across reported settings
Fetahi et al. (2025)	Albanian Facebook comments	XLM-RoBERTa + XAI	F1 up to 0.86

Note. Metrics originate from different datasets and are presented as published evidence; they are not directly comparable as one controlled experiment.

RESULTS AND DISCUSSION

The comparative evidence indicates that multilingual Transformer encoders provide the strong standard baseline for hate-speech detection when data span several languages and writing systems. XLM-R and mBERT repeatedly outperform classical machine-learning approaches because contextual embeddings preserve semantic relations that are lost in sparse lexical representations. Singh et al. (2023) reported macro-F1 values of 0.87, 0.83, and 0.77 for English, Hindi, and Hindi-English code-mixed settings using an mBERT-GRU framework, demonstrating the value of multilingual representations. Similar conclusions appear in the comparative analysis by Hashmi et al. (2024), where optimized transformer-based models improved multilingual detection across thirteen languages. These findings justify selecting XLM-RoBERTa as the backbone of the proposed optimized architecture, especially when a moderation system must operate across heterogeneous linguistic communities.

Cross-lingual transfer is important for languages with labeled hate-speech data. Bigoulaeva et al. (2023) showed that label modification and bootstrapping can improve zero-shot transfer by reducing mismatches between source and target annotation schemes. This issue is critical because hate-speech labels are not identical across datasets; one corpus may distinguish offensive language, while another focuses on protected-group targeting or severity. Firmino et al. (2024) likewise demonstrated that cross-lingual learning can produce substantial gains for Portuguese hate-speech detection by transferring knowledge from richer resources. Therefore, the optimized architecture should align label definitions before fine-tuning and should use a output ontology. Such alignment reduces negative transfer, improves comparability, and makes multilingual predictions more suitable for moderation policies that require consistent decisions across languages.

Code-mixed language remains a difficult condition because social-media users combine languages, abbreviations, transliteration, and nonstandard spelling within a

message. Biradar et al. (2022) demonstrated that transformer and translation-based strategies can improve detection for Hinglish, but translation may erase pragmatic cues and cultural insults. Kusuma and Chowanda (2023) showed that Indonesian hate-speech detection benefits from contextual IndoBERTweet representations combined with sequential modeling, confirming the importance of domain-specific language variation. The proposed optimization therefore favors joint multilingual encoding over mandatory translation, while retaining translation as an auxiliary augmentation strategy. Token-level normalization should preserve hashtags, mentions, emojis, and repeated characters instead of removing them indiscriminately. This design is expected to improve robustness for Indonesian-English mixtures and informal multilingual expressions common in social-media conversations.

Class imbalance affects hate-speech systems because hateful instances are less frequent than neutral or merely offensive content. Accuracy can remain high even when minority hate categories are poorly detected. Mnassri et al. (2024) addressed multilingual hate-speech detection with a semi-supervised generative adversarial approach, using unlabeled data and improving minority-class learning. Indonesian studies report uneven performance across hate and offensive categories, including Amalia and Suyanto (2024), whose BERT model achieved strong weighted performance but lower F1 for the hate class. Therefore, the optimized Transformer uses class-balanced focal loss rather than cross-entropy alone. The objective emphasizes minority examples, reduces domination by majority classes, and provides an optimization target for sensitive moderation tasks where false negatives can carry substantial consequences in practice today.

Functional robustness must be evaluated separately from held-out accuracy. Röttger et al. (2022) introduced Multilingual HateCheck with thirty-four functional categories across ten languages and showed that multilingual models still fail on negation, quotation, counter-speech, slur reclamation, and implicit targeting. This evidence suggests that train-test splits cannot reveal whether a detector has learned generalizable hate concepts or merely dataset-specific lexical shortcuts. The proposed evaluation framework therefore supplements macro-F1 with functionality-based stress tests. Each language should include contrastive cases that alter only one pragmatic feature, such as inserting negation, changing the target group, or converting a hateful statement into quoted condemnation. A model is considered optimized only when improvements persist under these perturbations, because practical moderation requires reliability beyond average benchmark scores.

Adversarial robustness is a requirement for deployment because users can evade moderation by replacing characters, inserting spaces, using homoglyphs, or substituting synonyms. Training multilingual Transformers with adversarial examples improves resistance to such manipulations. The study by Ryzhova et al. (2022) showed that adversarially enriched training and randomized ensembles reduced performance degradation across English, Hindi, and Russian hate-detection tasks. Accordingly, the proposed architecture incorporates embedding perturbation during fine-tuning and text-level augmentation for obfuscation patterns. This optimization is preferable to aggressive preprocessing, which may remove social-media signals. Robust training encourages the encoder to rely on contextual evidence rather than isolated trigger words. The expected outcome is more stable macro-F1 under adversarial edits without requiring a detector for every spelling variation encountered online.

Language-specific adaptation remains valuable when a multilingual backbone is used. Indonesian social media contains slang, abbreviations, identity terms, and pragmatic

conventions that may be underrepresented in multilingual pretraining. The IndoBERT study by Dharmawan et al. (2023) achieved high classification accuracy for Indonesian hate speech, while Nabiilah et al. (2023) found IndoBERT superior to multilingual alternatives for toxic Indonesian comments. Multilingual coverage alone does not guarantee optimal semantic sensitivity in every language. The proposed system applies language-adaptive fine-tuning before task-specific training, using unlabeled in-domain text when available. This stage updates representations toward local vocabulary without discarding cross-lingual knowledge. A multilingual classifier can then be supplemented by language adapters, allowing specialization while preserving one architecture and consistent decision boundaries across supported languages.

Explainability should accompany performance because hate-speech moderation can affect visibility, participation, and platform sanctions. Siddiqui et al. (2024) demonstrated multilingual hate-speech detection with explainable AI and Transformers, emphasizing the value of identifying influential features behind model decisions. Fetahi et al. (2025) similarly combined transformer models with explainability for a low-resource Albanian dataset and reported XLM-RoBERTa performance. In the proposed architecture, token attribution is used as a layer rather than as proof of causal reasoning. Explanations should help moderators detect spurious correlations, such as models treating identity terms themselves as hateful. Regular audits can compare attribution patterns across demographic groups and languages. This approach supports accountable deployment by linking model confidence, explanatory cues, and human review instead of enforcing uncertain predictions.

Sentiment information can support hate-speech detection, but it should function as an auxiliary signal rather than a substitute for hate labels. Negative sentiment includes criticism, frustration, and disagreement that may not attack protected groups, whereas hateful language can appear sarcastic or superficially positive. The optimized architecture therefore adopts multi-task learning: the Transformer encoder predicts both sentiment polarity and hate-speech labels, while the hate objective receives greater weight. This design encourages representation without collapsing concepts. Chiril et al. (2022) showed that emotionally informed modeling can improve multi-target hate-speech analysis, supporting the use of auxiliary features. In practice, sentiment-aware learning is expected to improve ambiguous cases when hostility is implicit, while calibration prevents negative but legitimate speech from being categorized as hateful.

Zero-shot and few-shot transfer are essential for extending moderation to languages that lack large annotated corpora. Zia et al. (2022) reported an average macro-F1 improvement of 7.6% through pseudo-label fine-tuning for zero-shot cross-lingual hate-speech detection across multiple languages and scripts. More recently, Hashmi et al. (2025) proposed cross-lingual meta-learning to improve adaptation to low-resource languages. These results support a two-stage optimization strategy: first train a shared multilingual representation on diverse source languages, then adapt selectively using a small number of target-language samples or high-confidence pseudo-labels. Confidence thresholds must be conservative because erroneous pseudo-labels can amplify bias. When carefully controlled, this procedure can reduce annotation requirements while maintaining language-specific sensitivity, making broader multilingual coverage technically and economically feasible for moderation systems.

User and conversational context can resolve cases that are ambiguous at the sentence level. Rehman et al. (2023) showed that multilingual abusive-content detection improves

when textual representations are combined with user history and post affinity, particularly for low-resource Indic languages and code-mixed text. contextual metadata introduces privacy and fairness concerns, so the proposed architecture treats context as bounded. Classifier relies on text, while conversational context can be invoked for uncertain predictions near the decision threshold. This selective design reduces data collection and limits the risk that demographic or behavioral proxies dominate classification. It supports deployment: platforms with privacy constraints can operate the text-only model, whereas moderated environments with contextual access can use a context encoder under governance and audit controls.

Calibration is necessary because Transformer probabilities are overconfident, after fine-tuning on imbalanced datasets. A moderation system should distinguish between a hateful prediction and an uncertain borderline case requiring human judgment. The proposed optimization applies temperature scaling on a calibration set and evaluates expected calibration error alongside macro-F1. This recommendation is consistent with the reliability concerns highlighted by Miok et al. (2022), who emphasized uncertainty-aware hate-speech detection for safer moderation decisions. Calibrated confidence enables risk-based routing: high-confidence cases can be prioritized for automated filtering, medium-confidence cases can receive human review, and low-confidence cases can be retained for active learning. Such a workflow converts model uncertainty into a resource, reducing false enforcement while generating informative examples for retraining and refinement across languages.

The multilingual studies indicate that optimization should integrate several complementary mechanisms rather than rely on a single architectural modification. UA-HSD-2025 experiments by Ahmad et al. (2025) showed that XLM-R achieved high performance across Arabic, Urdu, and multilingual settings, while translation-based alternatives varied by language. Cross-lingual contrastive and adversarial learning reported by Almahdi et al. (2025) further demonstrates that representation alignment is beneficial for low-resource transfer. Therefore, the proposed configuration combines XLM-RoBERTa, language-adaptive tuning, class-balanced focal loss, auxiliary sentiment learning, adversarial perturbation, and calibration. Each component addresses a failure mode: vocabulary mismatch, imbalance, affective ambiguity, evasion, or uncertainty. Ablation analysis should remove one component at a time to verify that complexity produces benefits rather than benchmark gains in multilingual deployments today.

Indonesian evidence reinforces the need for evaluation within a multilingual framework. Rokhim et al. (2024) combined IndoBERT with recurrent ensemble components for multi-level, multi-label hate-speech detection and reported improvements over individual models across categories. Herman and Romadhony (2024) used IndoBERT within an Indonesian political-hate context, illustrating how domain-specific discourse can change both vocabulary and error patterns. Consequently, the proposed system should not report only one global multilingual score. Results must be disaggregated by language, target category, severity, and domain so that weak groups remain visible. Macro-F1, per-class recall, and confusion matrices are more informative than accuracy alone. This reporting strategy prevents high-resource languages from masking failures in Indonesian or other lower-resource settings and supports data collection where errors are concentrated.

Overall, the evidence supports an optimized Transformer architecture as a foundation for multilingual hate-speech detection, but it shows that model quality depends on

evaluation design and governance. Ramos et al. (2024) concluded that the Transformer era has improved hate-speech detection while leaving unresolved problems involving dataset bias, generalization, explainability, and evolving language. The synthesis extends that conclusion by positioning optimization as a multi-objective process balancing macro-F1, cross-lingual transfer, functional robustness, calibration, and fairness. The resulting system should be deployed with human oversight, revalidation, and language-specific audits. Future implementation should prioritize public benchmark replication and prospective testing on collected data, because social-media vocabulary and political contexts change rapidly. Technical performance is necessary, but sustainable moderation requires transparent policies and empirical monitoring.

Table 2. Recommended optimization components and evaluation objectives

Component	Primary Purpose	Recommended Evaluation
Language-adaptive fine-tuning	Adapt multilingual representations to local slang and domain vocabulary	Language-wise macro-F1
Sentiment auxiliary head	Enrich affective representation without equating negativity with hate	Ablation on implicit hate and non-hateful criticism
Class-balanced focal loss	Improve minority hate-category learning	Per-class recall and macro-F1
Adversarial perturbation	Increase resistance to obfuscation and spelling attacks	Perturbation stress tests
Temperature scaling	Improve confidence reliability	Expected calibration error and human-review routing

CONCLUSION

This study concludes that optimized multilingual Transformer architectures provide a strong methodological foundation for hate-speech detection in heterogeneous social-media environments, but the most important gains arise from training and evaluation strategy rather than from encoder selection alone. Evidence from 2022–2025 consistently supports multilingual contextual representations, cross-lingual transfer, language adaptation, and class-sensitive optimization. XLM-RoBERTa is therefore proposed as the shared encoder, complemented by language-adaptive fine-tuning, focal loss, adversarial perturbation, and calibrated confidence. The literature also demonstrates that aggregate accuracy is insufficient because class imbalance and language dominance can conceal severe weaknesses in minority hate categories and low-resource languages.

The sentiment component should remain auxiliary. Affective information can enrich contextual representations and support recognition of implicit hostility, yet negative sentiment must not be equated with hate speech. The proposed multi-task architecture addresses this distinction through separate sentiment and hate-speech heads. In addition, functional tests, adversarial perturbations, and language-wise reporting should accompany

ordinary macro-F1 evaluation. Explainability and human review are especially important for uncertain predictions because moderation decisions can affect participation and expression. Consequently, an optimized system should be evaluated not only for predictive performance but also for robustness, calibration, consistency across languages, and the transparency of its failure modes.

Future research should implement the proposed architecture on harmonized public multilingual corpora and newly collected Indonesian social-media data using prospectively registered preprocessing and evaluation protocols. Ablation studies are needed to quantify the independent contribution of sentiment-aware learning, focal loss, adversarial training, language adapters, and probability calibration. Further work should also examine temporal drift, emerging slang, multimodal content, and fairness across identity groups. Such research would move multilingual hate-speech detection from benchmark-oriented optimization toward reliable operational moderation. The central implication is that technological improvement must remain tied to reproducibility and human governance: a model that is accurate on average but unstable across languages or uncertain contexts cannot be considered fully optimized for real social-media environments.

REFERENCES

1. Ahmad, M., Waqas, M., Hamza, A., Usman, S., Batyrshin, I., & Sidorov, G. (2025). UA-HSD-2025: Multi-Lingual Hate Speech Detection from Tweets Using Pre-Trained Transformers. Computers, 14(6), 239. https://doi.org/10.3390/computers14060239
2. Almahdi, A. J., Mohades, A., Akbari, M., & Heidary, S. (2025). Enhancing cross-lingual hate speech detection through contrastive and adversarial learning. Engineering Applications of Artificial Intelligence, 147, 110296. https://doi.org/10.1016/j.engappai.2025.110296
3. Amalia, F. S., & Suyanto, Y. (2024). Offensive Language and Hate Speech Detection using BERT Model. IJCCS (Indonesian Journal of Computing and Cybernetics Systems). https://doi.org/10.22146/ijccs.99841
4. Bigoulaeva, I., Hangya, V., Gurevych, I., & Fraser, A. (2023). Label modification and bootstrapping for zero-shot cross-lingual hate speech detection. Language Resources and Evaluation, 57, 1515–1546. https://doi.org/10.1007/s10579-023-09637-4
5. Biradar, S., Saumya, S., & Chauhan, A. (2022). Fighting hate speech from bilingual Hinglish speaker's perspective, a transformer- and translation-based approach. Social Network Analysis and Mining, 12, 87. https://doi.org/10.1007/s13278-022-00920-w
6. Chiril, P., Pamungkas, E. W., Benamara, F., Moriceau, V., & Patti, V. (2022). Emotionally Informed Hate Speech Detection: A Multi-target Perspective. Cognitive Computation, 14, 322–352. https://doi.org/10.1007/s12559-021-09862-5
7. Dharmawan, S., Mawardi, V. C., & Perdana, N. J. (2023). Klasifikasi Ujaran Kebencian Menggunakan Metode FeedForward Neural Network (IndoBERT). Jurnal Ilmu Komputer dan Sistem Informasi, 11(1). https://doi.org/10.24912/jiksi.v11i1.24066
8. Fetahi, E., Susuri, A., Hamiti, M., Kastrati, Z., Canhasi, E., & Misini, A. (2025). Enhancing social media hate speech detection in low-resource languages using

transformers and explainable AI. Social Network Analysis and Mining, 15, 82. https://doi.org/10.1007/s13278-025-01497-w
9. Firmino, A. A., Baptista, C. S., & de Paiva, A. C. (2024). Improving hate speech detection using cross-lingual learning. Expert Systems with Applications, 235, 121115. https://doi.org/10.1016/j.eswa.2023.121115
10. Hashmi, E., Yayilgan, S. Y., Hameed, I. A., Yamin, M. M., Ullah, M., & Abomhara, M. (2024). Enhancing Multilingual Hate Speech Detection: From Language-Specific Insights to Cross-Linguistic Integration. IEEE Access, 12, 121507–121537. https://doi.org/10.1109/ACCESS.2024.3452987
11. Hashmi, E., Yayilgan, S. Y., & Abomhara, M. (2025). Metalinguist: enhancing hate speech detection with cross-lingual meta-learning. Complex & Intelligent Systems, 11, 179. https://doi.org/10.1007/s40747-025-01808-w
12. Herman, F. R., & Romadhony, A. (2024). Buzzer Account Detection in Political Hate Tweets Using IndoBERT and Ensemble Learning: Case Study of the Indonesian Presidential Election 2024. International Journal on ICT, 10(2), 184–194. https://doi.org/10.21108/ijoict.v10i2.1012
13. Kusuma, J., & Chowanda, A. (2023). Indonesian Hate Speech Detection Using IndoBERTweet and BiLSTM on Twitter. JOIV: International Journal on Informatics Visualization, 7(3), 773–780. https://doi.org/10.30630/joiv.7.3.1035
14. Liu, L., Xu, D., Zhao, P., Zeng, D. D., Hu, P. J.-H., Zhang, Q., Luo, Y., & Cao, Z. (2023). A cross-lingual transfer learning method for online COVID-19-related hate speech detection. Expert Systems with Applications, 234, 121031. https://doi.org/10.1016/j.eswa.2023.121031
15. Miok, K., Škrlj, B., Zaharie, D., & Robnik-Šikonja, M. (2022). To BAN or Not to BAN: Bayesian Attention Networks for Reliable Hate Speech Detection. Cognitive Computation, 14, 353–371. https://doi.org/10.1007/s12559-021-09826-9
16. Mnassri, K., Farahbakhsh, R., & Crespi, N. (2024). Multilingual Hate Speech Detection: A Semi-Supervised Generative Adversarial Approach. Entropy, 26(4), 344. https://doi.org/10.3390/e26040344
17. Nabiilah, G. Z., Prasetyo, S. Y., Izdihar, Z. N., & Girsang, A. S. (2023). BERT base model for toxic comment analysis on Indonesian social media. Procedia Computer Science, 216, 714–721. https://doi.org/10.1016/j.procs.2022.12.188
18. Ramos, G., Batista, F., Ribeiro, R., Fialho, P., Moro, S., Fonseca, A., Guerra, R., Carvalho, P., Marques, C., & Silva, C. (2024). A comprehensive review on automatic hate speech detection in the age of the transformer. Social Network Analysis and Mining, 14, 204. https://doi.org/10.1007/s13278-024-01361-3
19. Rehman, M. Z. U., Mehta, S., Singh, K., Kaushik, K., & Kumar, N. (2023). User-aware multilingual abusive content detection in social media. Information Processing & Management, 60(5), 103450. https://doi.org/10.1016/j.ipm.2023.103450
20. Rokhim, I. F., Sarno, R., Septiyanto, A. F., Haryono, A. T., & Sabilla, S. I. (2024). IndoBERT-Based Ensemble Learning for Multi-Level Multi-Label Hate Speech Detection in Indonesian Social Media. In 2024 Beyond Technology Summit on Informatics International Conference (BTS-I2C) (pp. 456–461). IEEE. https://doi.org/10.1109/BTS-I2C63534.2024.10942204
21. Röttger, P., Seelawi, H., Nozza, D., Talat, Z., & Vidgen, B. (2022). Multilingual HateCheck: Functional Tests for Multilingual Hate Speech Detection Models. In

Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH) (pp. 154–169). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.woah-1.15

22. Ryzhova, A., Devyatkin, D., Volkov, S., & Budzko, V. (2022). Training Multilingual and Adversarial Attack-Robust Models for Hate Detection on Social Media. Procedia Computer Science, 213, 196–202. https://doi.org/10.1016/j.procs.2022.11.056
23. Sahir, S. H. (2022). Metodologi Penelitian. KBM Indonesia.
24. Siddiqui, J. A., Yuhaniz, S. S., Shaikh, G. M., Soomro, S. A., & Mahar, Z. A. (2024). Fine-Grained Multilingual Hate Speech Detection Using Explainable AI and Transformers. IEEE Access, 12, 143177–143192. https://doi.org/10.1109/ACCESS.2024.3470901
25. Singh, P., Singh, N. K., Monika, & Chand, S. (2023). mBERT-GRU multilingual deep learning framework for hate speech detection in social media. Journal of Intelligent & Fuzzy Systems, 44(5), 8177–8192. https://doi.org/10.3233/JIFS-222057
26. Sugiyono. (2022). Metode Penelitian Kuantitatif, Kualitatif, dan R&D. Alfabeta.
27. Zia, H. B., Castro, I., Zubiaga, A., & Tyson, G. (2022). Improving Zero-Shot Cross-Lingual Hate Speech Detection with Pseudo-Label Fine-Tuning of Transformer Language Models. Proceedings of the International AAAI Conference on Web and Social Media, 16(1), 1435–1439. https://doi.org/10.1609/icwsm.v16i1.19402