



An Applied Evaluation of Multi-Label Hate Speech Detection for Indonesian Digital Platforms

Dwijja Wisnu Brata^{1,2}, Arif Djunaidy^{3*}, Daniel Siahaan¹, Edio da Costa⁴

¹Department of Informatics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

²Department of Information System, Brawijaya University, Malang, Indonesia

³Department of Information Systems, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

⁴Department of Computer Science, School of Engineering and Science, Dili Institute of Technology, Dili, Timor-Leste

*arif.djunaidy@its.ac.id

Abstract

The rapid growth of user-generated content on digital platforms has increased the difficulty of moderating hate-related expressions, particularly in linguistically diverse environments such as Indonesian social media. Automated hate speech detection systems are therefore expected to operate not only with reliable predictive behavior but also with practical efficiency under large-scale deployment conditions. This study reports an applied evaluation of Transformer-based models for multi-label hate speech detection on Indonesian digital platforms. Rather than introducing a new classification architecture, the work focuses on assessing multiple pretrained language models within a unified and reproducible evaluation framework. The analysis examines overall model behavior, per-label performance tendencies, inference efficiency, and common error patterns under realistic multi-label settings. The results indicate that IndoBERT-based models (indobenchmark/indobert-base-p1 and cahya/bert-base-indonesian) achieved the strongest predictive performance for multi-label hate speech detection, although performance differences across the evaluated Transformer models remained relatively incremental. Experimental results show that the best-performing model achieved a macro-F1 score of 0.9742 and a micro-F1 score of 0.9749, while other Transformer models demonstrated competitive performance with macro-F1 values ranging from 0.955 to 0.964. In terms of efficiency, distilled models provided faster inference (approximately 5–7 ms per sample) compared to full-size models (8–11 ms per sample), highlighting a practical trade-off between predictive performance and computational cost. These findings emphasize the importance of practical evaluation strategies and suggest that flexible model configurations are more suitable than reliance on a single high-capacity model in real-world moderation systems.

Keywords: hate speech; indonesian text; multi-label; text filtering; transformers

How to Cite: D. W. Brata, A. Djunaidy, D. O. Siahaan, and Edio da Costa, "An Applied Evaluation of Multi-Label Hate Speech Detection for Indonesian Digital Platforms", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 10, no. 3, pp. 834 - 842, Jun. 2026.
Permalink/DOI: <https://doi.org/10.29207/resti.v10i3.7591>

Received: February 19, 2026

Accepted: May 23, 2026

Available Online: June 28, 2026

*This is an open-access article under the [CC BY 4.0 License](#)
Published by [Ikatan Ahli Informatika Indonesia](#)*

1. Introduction

The rapid expansion of digital communication platforms has fundamentally transformed how information is produced, disseminated, and consumed. Alongside their social and economic benefits, these platforms have also intensified the spread of hate-related content, posing persistent challenges to content governance and social cohesion, particularly within the Indonesian digital environment [1]. The scale and immediacy of user-generated content, such as hate speech, require moderation mechanisms that can operate continuously under high-volume conditions without sacrificing reliability. Research on hate speech detection over the past five years has shown a

significant shift from binary classification approaches toward multi-level and multi-label methods that are better able to capture the complexity of hateful expressions.

Empirical analyses of online discourse indicate that hate speech and abusive expressions remain prevalent on social media platforms despite the implementation of platform-level policies and automated filtering mechanisms [2]. This persistence suggests that moderation challenges are not solely the result of insufficient regulation or enforcement. Instead, they are also shaped by technical limitations in detection systems that must handle informal language, evolving expressions, and diverse communicative contexts.

From a broader research perspective, automated hate speech detection has been widely studied using natural language processing techniques. Survey-based analyses report steady methodological progress, while also emphasizing that hate speech detection remains inherently difficult due to linguistic variability, contextual dependency, and annotation subjectivity across datasets and platforms [3]. These characteristics make generalization across domains and languages particularly challenging.

From an engineering standpoint, manual moderation is no longer feasible for large-scale digital platforms, leading to increased reliance on automated content moderation systems [4]. Such systems are expected to process large volumes of content under strict latency and throughput constraints. Consequently, moderation solutions must be designed with equal attention to detection quality and operational efficiency.

In operational settings, hate-related expressions rarely conform to a single, clearly defined category. Abusive content often spans multiple dimensions, including target type, contextual framing, and degrees of severity, which complicates straightforward classification strategies [5]. This multi-dimensional nature introduces ambiguity that is difficult to capture using simplified modeling assumptions.

Additional limitations arise from the long-tail distribution of hate speech categories in real-world data. Empirical findings show that infrequent and context-dependent forms of hate are consistently harder to detect, even when overall performance metrics appear strong [6]. This imbalance-driven behavior reduces robustness and highlights the need for evaluation settings that better reflect realistic moderation scenarios.

Recent advances in Transformer-based language models have led to substantial improvements in text classification tasks, including hate speech and abusive language detection [7]. In the Indonesian language context, both language-specific and multilingual pretrained models have demonstrated strong representational capability across a range of downstream natural language processing applications [8]. These developments have encouraged the adoption of Transformer-based architectures in moderation-related research.

However, much of the existing literature continues to prioritize predictive accuracy as the primary evaluation objective. System-level considerations such as inference efficiency, training stability, and computational cost are often treated as secondary factors, despite their importance in real-world deployment environments [9]. In large-scale moderation systems, these factors directly influence throughput, scalability, and long-term operational viability.

In practice, performance differences among high-capacity language models are frequently modest,

whereas variations in inference latency and resource consumption can have a substantial impact on system behavior [10]. Small differences in processing time may accumulate significantly when models are deployed at scale. This makes efficiency-aware evaluation essential from an applied systems perspective.

Despite recent advances in Transformer-based hate speech detection, comprehensive evaluations that jointly examine predictive performance and computational efficiency under consistent experimental conditions remain limited, particularly for multi-label hate speech detection in Indonesian-language data [11]. This limitation restricts the ability to derive system-level insights that are directly applicable to real-world deployment. Consequently, model selection in existing studies is often driven primarily by benchmark accuracy, with insufficient consideration of operational constraints such as inference latency and resource efficiency.

Furthermore, the practical integration of models with varying capacity and efficiency characteristics into moderation pipelines has received relatively little attention. Prior studies have shown that implicit and explicit forms of abusive language exhibit distinct detection patterns, which introduce additional complexity in system design for real-world applications [12]. Without a systematic evaluation of these factors, deployed moderation systems may fail to align with actual usage requirements and operational conditions.

In response to these challenges, this study presents an applied evaluation of Transformer-based models for multi-label hate speech detection on Indonesian digital platforms. Existing works have demonstrated the effectiveness of Transformer architectures in abusive language detection. For example, Mozafari et al. [7] showed that BERT-based transfer learning significantly improves performance in multilingual settings, while Vidgen and Derczynski [2] highlighted the critical role of dataset quality and annotation strategies. In the Indonesian context, the dataset introduced by Ibrohim and Budi [13] has become a widely adopted benchmark for multi-label hate speech detection.

However, despite these developments, most prior studies predominantly emphasize classification accuracy, with limited attention to system-level factors such as inference efficiency, runtime latency, and deployment feasibility. In large-scale moderation systems, these factors are equally critical as predictive performance, as they directly affect scalability and operational viability. Therefore, a systematic evaluation that jointly considers predictive behavior and efficiency characteristics is necessary to support practical deployment decisions.

Rather than proposing a new classification architecture, this study evaluates multiple pretrained Transformer-based language models within a unified and controlled experimental framework. The analysis focuses on system-relevant aspects, including per-label

performance, inference efficiency, and accuracy-latency trade-offs, with the goal of supporting informed model selection and deployment strategies in large-scale moderation environments.

2. Methods

2.1 Framework Overview

This study adopts an applied intelligent framework to evaluate multi-label hate speech detection models under consistent and deployment-oriented settings. Rather than introducing a new classification architecture, the framework is designed to enable systematic and reproducible comparison across multiple Transformer-based language models. The emphasis is placed on system-level evaluation, including consistency,

operational relevance, and deployment feasibility. This design ensures comparability across models by maintaining consistent preprocessing, training, and evaluation procedures. Consequently, the evaluation outcomes reflect system-level behavior and deployment-oriented characteristics rather than variations introduced by experimental configuration.

The overall framework pipeline is illustrated in Figure 1, which depicts the end-to-end processing stages from raw text input to multi-label prediction outputs. The pipeline consists of four main components: dataset preparation, text preprocessing, model fine-tuning, and evaluation. This structure follows common practices in Transformer-based natural language processing, where pretrained models are adapted to downstream tasks through controlled fine-tuning procedures [14].

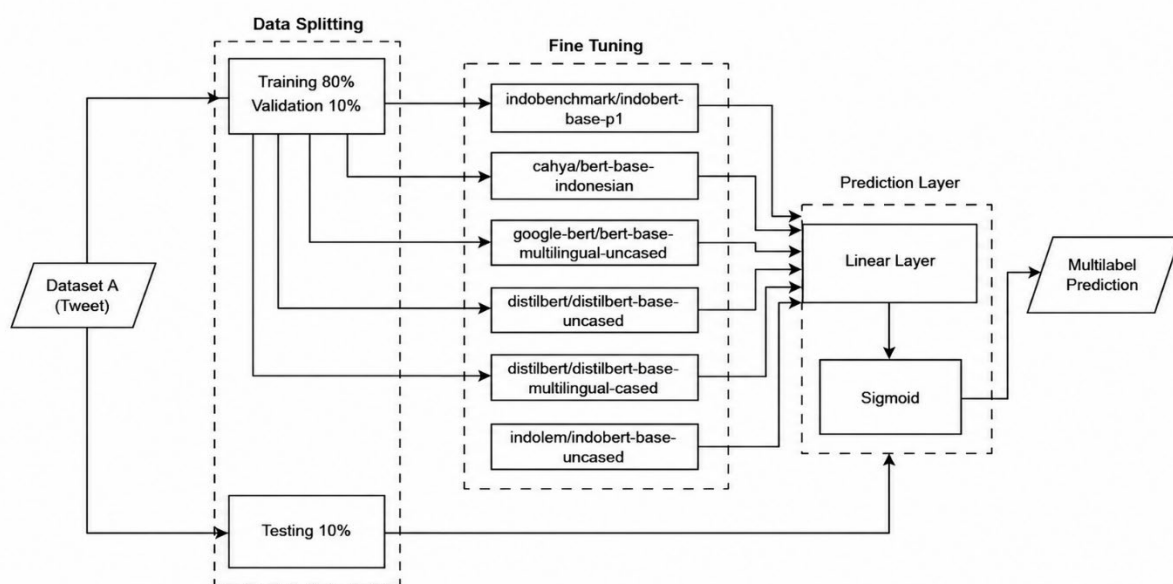


Figure 1. Applied Intelligent Framework for Multi-Label Hate Speech Detection

To support multi-label learning under realistic moderation conditions, the framework incorporates label-aware training and threshold-based prediction strategies. Class imbalance is addressed during training through weighted loss functions, while decision thresholds are applied at the prediction stage to accommodate overlapping labels. These design choices ensure that the evaluation setting aligns with operational moderation scenarios rather than simplified benchmark conditions.

2.2 Dataset Preparation

The dataset used in this study consists of 52,252 Indonesian-language social media posts collected from Twitter (now known as X). The corpus was constructed by combining the publicly available Multi-Label Indonesian Twitter Hate Speech dataset developed by Ibrohim and Budi [13] with additional tweets obtained through further crawling. The initial dataset, retrieved from the UCI Machine Learning Repository[15], contained 13,169 Indonesian-language tweets annotated for hate speech and abusive content. In this

study, the dataset was expanded with 39,083 newly collected tweets, resulting in a total of 52,252 instances. The additional tweets were gathered through targeted crawling of public Twitter conversations, followed by manual filtering to ensure contextual relevance (Indonesian language, Twitter discourse context, and removal of duplicates). Manual annotation was conducted using the same multi-label schema as Ibrohim and Budi [13] to maintain consistency. Each instance may be associated with more than one label, reflecting the overlapping and multifaceted nature of abusive expressions commonly observed in online platforms, and this multi-label structure allows a single post to represent compound moderation violations.

The definitions of the hate speech categories and their corresponding data distributions are summarized in Table 1. The label set captures multiple dimensions of abusive content, including general hate indicators, target types, and severity levels. This structure supports more realistic modeling of moderation requirements compared to single-label formulations.

Table 1. Definition of Multi-Label Hate Speech Categories

Label	Label Frequency (%)	Description
HS	42.23	General hate or abusive language
Abusive	38.29	Offensive but non-hate expression
HS_Individual	27.15	Targets specific individuals
HS_Group	15.08	Target groups or communities
HS_Religion	6.02	Based on religious identity
HS_Race	4.3	Based on race or ethnicity
HS_Physical	2.45	Refers to physical traits or disabilities
HS_Gender	2.32	Based on gender or sexual identity
HS_Other	28.4	Ambiguous or mixed expressions
HS_Weak	25.69	Mild hate intensity

The dataset used in this study is derived from the Multi-label Indonesian Twitter Hate Speech Dataset available through the UCI Machine Learning Repository. The data are partitioned into training, validation, and testing subsets using an 80:10:10 ratio to ensure sufficient representation for model learning, intermediate validation, and final evaluation [11]. The validation set is used to monitor model performance during training and to maintain consistency across experiments, while the testing set is reserved for the final evaluation of model generalization [11]. All models are trained using the same optimization settings to ensure a fair and controlled comparison. Under this experimental design, observed performance differences can be attributed primarily to model architecture and representational capacity, rather than variations introduced by model-specific tuning procedures [11]. To prevent information leakage, duplicate entries are removed independently within each subset after the data splitting process, ensuring that identical samples do not appear across training and testing data. This approach follows established practices in supervised multi-label learning and supports reliable and unbiased evaluation [11].

2.3 Text Preprocessing

Text preprocessing is conducted using standard procedures commonly applied in Transformer-based text classification tasks. The preprocessing pipeline includes text normalization, case folding, and removal of non-informative symbols, followed by subword tokenization aligned with each pretrained language model. These steps aim to reduce noise while preserving the semantic structure of the original text.

URLs, user mentions, and platform-specific tokens are normalized to generic placeholders to improve robustness across diverse input patterns [16]. No explicit data augmentation or synthetic oversampling techniques are applied at the preprocessing stage. This decision avoids introducing artificial patterns that could distort the original linguistic characteristics of the dataset [16].

Instead of modifying the data distribution, considerations related to class imbalance are addressed during training and evaluation. This approach preserves annotation consistency and supports analysis under conditions that closely resemble real-world moderation data, where imbalance is inherent and unavoidable [11].

2.4 Model Configuration and Fine-Tuning

Several pretrained Transformer-based language models are evaluated within the proposed framework to examine trade-offs between representational capacity and computational efficiency. All models are fine-tuned under identical training configurations to ensure fair and reproducible comparison across architectures. The use of uniform optimization settings allows observed differences in performance and efficiency to be attributed to model characteristics rather than experimental variation.

The evaluated models and their corresponding training configurations are summarized in Table 2. The model set includes both multilingual Transformer architectures and Indonesian-specific pretrained models to reflect realistic deployment options in heterogeneous moderation environments. Fine-tuning follows standard practices for adapting pretrained Transformer models to downstream classification tasks, including supervised optimization of task-specific output layers [17].

Table 2. Summary of Transformer-based language models

Model	Description
indobenchmark/indobert-base-pl	Indonesian BERT model trained on Wikipedia and news corpora using SentencePiece tokenizer [8]
indolem/indobert-base-uncased	IndoBERT model further pretrained and evaluated using the IndoLEM benchmark [14]
cahya/bert-base-indonesian	Indonesian BERT trained on large web corpora [16]
google-bert/bert-base-multilingual-uncased	Multilingual model covering 104 languages [15]
distilbert-base-uncased	Distilled version of BERT for faster inference [10]
distilbert-base-multilingual-cased	Compact multilingual model preserving capitalization [10]

Training is performed using a binary cross-entropy loss function adapted for multi-label classification, which enables independent optimization of overlapping labels while maintaining stable convergence behavior. In addition to multilingual representations, the framework incorporates Indonesian-specific pretrained models that are trained on large-scale local corpora and benchmarked for Indonesian natural language understanding tasks [18]. The inclusion of such models aims to better capture linguistic characteristics of Indonesian online discourse, including informal and context-dependent expressions commonly found in social media data [19].

To ensure consistent and reproducible training across all evaluated models, a fixed set of hyperparameters is applied during the fine-tuning process. These hyperparameters are selected based on commonly

adopted practices in Transformer-based text classification tasks and are kept identical across all models to ensure a fair and controlled comparison. This configuration is designed to balance training stability, computational efficiency, and generalization capability, while minimizing variability introduced by model-specific tuning.

All models are fine-tuned using the AdamW optimizer with a learning rate of 2×10^{-5} . Batch sizes of 16 and 32 are employed to analyze the trade-off between computational efficiency and gradient stability. The maximum sequence length is varied across 64, 128, and 256 tokens to capture different levels of contextual information. The models are trained for a fixed number of epochs under identical training conditions. Regularization techniques, including dropout (0.1) and weight decay (0.1), are applied to reduce overfitting and improve model generalization. Binary cross-entropy loss is used to support multi-label classification, with a sigmoid activation function in the output layer. A fixed decision threshold of 0.5 is applied to convert predicted probabilities into label assignments.

No extensive hyperparameter tuning is performed in this study, as the primary objective is to evaluate model behavior under consistent experimental conditions. This design ensures that observed performance differences are primarily influenced by model architecture rather than optimization strategies.

2.5 Evaluation

Model performance is evaluated using the macro-averaged F1 score as the primary metric. This metric assigns equal importance to each label and is particularly suitable for imbalanced multi-label classification problems, where rare categories may otherwise be overshadowed by dominant labels [11]. Macro-level evaluation supports analysis of system behavior across diverse moderation criteria. The validation set is used during training to monitor model performance and ensure consistent learning behavior across all evaluated models, while the testing set is reserved exclusively for final evaluation.

Micro-averaged F1, precision, and recall are also computed to provide complementary perspectives on overall performance. These metrics reflect aggregate system behavior and are commonly used in operational moderation settings to assess throughput-oriented effectiveness. Together, the selected metrics support balanced evaluation of predictive performance and robustness. No performance results or visualizations are presented in this section. All empirical findings are reported in the Results and Discussion section to maintain a clear separation between methodology and evaluation outcomes.

2.6 Implementation Details

All experiments are implemented using PyTorch version 2.0.1 in conjunction with the Hugging Face Transformers library version 4.42 [20]. This software

stack supports efficient fine-tuning and evaluation of Transformer-based language models and is widely adopted in applied machine learning research. Training and evaluation are conducted in a cloud-based GPU environment equipped with an NVIDIA T4 GPU with 16 GB of VRAM. Hardware configuration is kept consistent across experiments to ensure comparability of runtime-related observations. Unless explicitly stated otherwise, all hyperparameter values remain fixed across models to isolate the effects of model architecture on observed performance. The validation subset is used during training for intermediate performance monitoring and training stability assessment.

3. Results and Discussions

3.1 Overall Multi-Label Classification Performance

The overall multi-label classification performance of the evaluated Transformer-based models is summarized in Table 3. The table reports macro-averaged and micro-averaged F1 scores under a standardized experimental configuration. These metrics provide a balanced view of predictive effectiveness and system-level behavior.

Table 3. Multi-Label Classification Performance of Transformer-Based Models

Model	Macro-F1	Micro-F1
indobenchmark/indobert-base-pl	0.9742	0.9749
indolem/indobert-base-uncased	0.9562	0.9591
cahya/bert-base-indonesian	0.9715	0.9719
google-bert/bert-base-multilingual-uncased	0.9551	0.9551
distilbert-base-uncased	0.9590	0.9591
distilbert-base-multilingual-cased	0.9640	0.9645

Across all evaluated models, micro-average F1 scores are consistently higher than macro-averaged scores. This pattern reflects the influence of dominant labels in the dataset and is commonly observed in imbalanced multi-label classification settings. Macro-averaged metrics therefore offer a more informative indication of model behavior on less frequent hate speech categories, which are often more relevant from a moderation perspective [21].

Performance differences among the evaluated models remain relatively modest. Language-specific models pretrained on Indonesian corpora tend to achieve slightly higher macro-level scores, but the margins are limited. This observation aligns with prior findings that improvements in aggregate performance metrics do not necessarily translate into substantial gains in robustness across all hate speech categories [22].

From an applied system perspective, the results in Table 3 indicate that reliable multi-label hate speech detection can be achieved using multiple Transformer configurations. Rather than identifying a single dominant model, the findings suggest that system

designers may prioritize operational constraints, such as latency and resource availability, without substantially compromising predictive performance.

To further assess the robustness of the evaluated models, the potential risk of overfitting is examined through performance consistency on a held-out testing dataset. As shown in Table 3, the macro-F1 scores across different Transformer architectures remain within a relatively narrow range, indicating stable generalization behavior under the applied experimental setting. The relatively narrow performance range across evaluated models suggests comparable generalization behavior under the applied experimental setting.

In addition, regularization techniques such as dropout (0.1) and weight decay (0.1) are incorporated during training to improve generalization and reduce overfitting risk. It is important to note that, while these observations indicate stable generalization, the analysis is limited because evaluation was conducted using a fixed train-validation-test split. Future work may incorporate cross-validation or additional validation strategies to provide a more comprehensive assessment of model generalization. Nevertheless, under the controlled experimental setting adopted in this study, the results suggest a balanced trade-off between model complexity and generalization capability.

3.2 Per-Label Performance Analysis

The per-label performance of the evaluated models is summarized in Table 4, which reports F1 scores for each hate speech category under the multi-label evaluation setting. This analysis provides insight into how predictive behavior varies across different target types and severity levels. Such variation is expected in moderation tasks where labels differ in frequency, linguistic realization, and contextual dependency.

Table 4. F1-Score Distribution Across Hate Speech Categories

Label Category	F1-score
Hate Speech (HS)	0.993
Abusive	0.991
Target: Individual (HS_Individual)	0.991
Target: Group (HS_Group)	0.990
Target: Religion (HS_Religion)	0.994
Target: Race (HS_Race)	0.992
Target: Gender (HS_Gender)	0.993
Target: Physical (HS_Physical)	0.981
Other / Ambiguous (HS_Other)	0.883
Severity: Weak (HS_Weak)	0.945
Severity: Moderate (HS_Moderate)	0.942
Severity: Strong (HS_Strong)	0.982

Labels associated with explicit abusive expressions and frequently occurring targets exhibit consistently higher F1 scores across models. These categories typically contain clearer lexical and syntactic cues, making them easier to learn under supervised training. Similar observations have been reported in prior studies on hate speech and abusive language detection, where explicit forms of abuse are generally more reliably detected than implicit ones [21].

In contrast, labels related to implicit hate, ambiguous targeting, and lower-frequency categories tend to show reduced recall. These cases often rely on contextual inference, sarcasm, or culturally specific references rather than overt offensive terms. Previous empirical analyses have highlighted that such implicit expressions remain challenging for text-only classification models, particularly in imbalanced and noisy social media data [22].

Another factor contributing to per-label variation is the presence of overlapping annotations within the multi-label scheme. Individual instances may simultaneously express multiple targets or severity levels, increasing classification complexity. Models may correctly identify one dimension while failing to capture others, leading to partial predictions that are penalized in per-label evaluation, a behavior also observed in large-scale abusive language datasets [23].

From an applied moderation perspective, these per-label performance patterns suggest that automated systems are more reliable for detecting overt and commonly expressed forms of abuse. Conversely, nuanced cases involving implicit intent or overlapping violations require additional handling mechanisms beyond standalone model predictions. Prior work indicates that such limitations are more effectively addressed through system-level strategies, including confidence-based filtering and selective human review, rather than purely architectural modifications [24].

3.3 Model Comparison from an Efficiency Perspective

The trade-off between classification performance and inference efficiency across the evaluated models is illustrated in Figure 2. The figure relates macro-averaged F1 scores to average inference latency measured under a consistent experimental setup. This representation provides a system-oriented view of model suitability for deployment under different operational constraints.

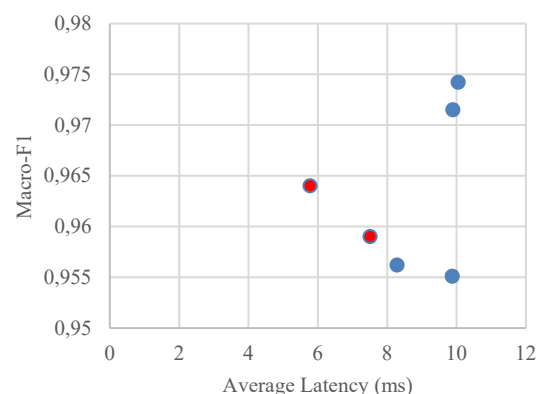


Figure 2. Trade-off between classification performance and inference latency

As shown in Figure 2, full-size Transformer models, indicated by blue markers, generally achieve slightly higher macro-level performance at the cost of increased inference latency. These models include Indonesian-

specific BERT variants and the multilingual BERT-base model, each containing approximately 110 million parameters. Similar trade-offs between representational capacity and computational overhead have been widely reported in studies examining Transformer deployment in resource-constrained environments [25].

In contrast, distilled Transformer variants, represented by red markers, demonstrate noticeably lower inference latency while maintaining competitive macro-average performance. By reducing the parameter count to approximately 66 million through knowledge distillation, these models achieve faster inference with limited degradation in classification effectiveness. Prior work has shown that such compression techniques are particularly beneficial for large-scale text processing systems where throughput and responsiveness are critical [26].

A closer inspection of runtime behavior reveals systematic latency differences among individual models. IndoBERT processes inputs at an average latency of approximately 10.05 ms, while the IndoLEM IndoBERT and Cahya Indonesian BERT models exhibit average latencies of around 8.29 ms and 9.90 ms, respectively. The multilingual BERT-based model operates at a comparable latency of approximately 9.88 ms, whereas DistilBERT and multilingual DistilBERT achieve faster inference times of approximately 7.51 ms and 5.78 ms, respectively.

From an applied systems perspective, these efficiency–performance trade-offs support flexible deployment strategies rather than a single optimal configuration. Lightweight distilled models are well suited for latency-sensitive or resource-constrained environments, while full-size language-specific models remain appropriate for secondary analysis of ambiguous or high-risk content. Such tiered or cascading moderation architectures have been recommended in prior work as an effective means of balancing throughput, accuracy, and operational cost in real-world moderation systems [27].

3.4 Error Patterns and Prediction Behavior

An analysis of prediction errors reveals that misclassifications are primarily driven by linguistic ambiguity and the inherent complexity of multi-label hate speech expressions. Unlike aggregate metrics, error-level inspection highlights how models behave when confronted with subtle, context-dependent language commonly found in social media discourse. These observations provide insight into practical limitations that are not immediately visible from overall performance scores.

One prominent source of error arises from implicit or indirect expressions of hate. Content that conveys derogatory intent through sarcasm, metaphor, or coded language frequently lacks explicit lexical cues, leading to under-prediction of hate-related labels. Similar challenges have been reported in large-scale analyses of

abusive behavior, where implicit hate consistently proves more difficult to detect than overt forms [28].

Another recurring error pattern is associated with overlapping labels within the multi-label annotation scheme. Posts that simultaneously reference multiple targets or combine target and severity dimensions may result in partial predictions, where only a subset of relevant labels is identified. Prior work has shown that such behavior is common in datasets with fine-grained annotation schemes and reflects limitations in capturing complex relational semantics using text-only models [29].

False positive predictions are most often observed in content that exhibit strong sentiment without explicit hateful intent. Heated political discussions or emotionally charged personal opinions may trigger hate-related predictions due to surface-level intensity rather than genuine abuse. Studies on offensive language detection have similarly noted that heightened sensitivity can increase false positives, reinforcing the need for confidence-aware thresholds and selective human review in applied moderation systems [30].

3.5 Discussion Summary and Practical Implications

This study provides an applied evaluation of Transformer-based models for multi-label hate speech detection on Indonesian digital platforms, with emphasis on system-level behavior rather than architectural novelty. Across overall performance, per-label analysis, efficiency trade-offs, and error patterns, the results demonstrate that reliable detection can be achieved using multiple model configurations under consistent evaluation settings. The findings highlight that differences among models are generally incremental and must be interpreted in relation to operational requirements.

From a performance perspective, the analysis shows that strong aggregate metrics do not necessarily imply uniform robustness across all hate speech categories. Explicit and frequently occurring forms of abuse are detected more reliably, while implicit, context-dependent, and overlapping violations remain challenging. These patterns reflect realistic moderation conditions and underscore the importance of per-label analysis when assessing system readiness for deployment.

The efficiency-oriented evaluation further emphasizes that deployment suitability depends not only on predictive accuracy but also on inference latency and computational cost. Distilled models offer practical advantages for high-throughput or latency-sensitive environments, while full-size language-specific models remain relevant for scenarios requiring deeper semantic analysis. This trade-off supports flexible deployment strategies in which model selection is guided by system constraints rather than marginal performance differences alone.

Taken together, the results suggest that effective hate speech moderation is best approached as a system design problem rather than a single-model optimization task. Practical moderation systems benefit from modular and tiered architectures that combine automated detection with confidence-aware decision mechanisms and selective human oversight. These insights provide a system-oriented foundation for designing scalable and deployment-ready hate speech detection frameworks for Indonesian digital platforms and naturally motivate the conclusions discussed in the following section.

4. Conclusions

This study conducted an applied evaluation of Transformer-based models for multi-label hate speech detection on Indonesian digital platforms, with emphasis on system-level performance and deployment readiness. By using a unified experimental framework, the evaluation examined overall predictive behavior, per-label performance characteristics, inference efficiency, and error patterns under realistic multi-label conditions.

The results show that IndoBERT-based models (indobenchmark/indobert-base-pl and cahya/bert-base-indonesian) provide the most stable performance for multi-label hate speech detection on Indonesian digital platforms, while other Transformer configurations are also capable of supporting reliable detection. Performance differences across models are generally incremental, while efficiency-related factors such as inference latency introduce more pronounced practical implications. Language-specific models demonstrate strong representation of Indonesian hate speech patterns, whereas distilled models provide competitive performance with substantially improved efficiency, making them suitable for high-throughput moderation scenarios.

From a system design perspective, these findings indicate that effective hate speech moderation does not depend on a single high-capacity model. Instead, deployment-ready moderation systems benefit from flexible configurations in which model selection is guided by operational constraints and moderation objectives. Overall, this work provides applied insights that support the design of scalable, efficient, and practical hate speech detection systems for Indonesian digital platforms.

Acknowledgements

The authors declare no conflict of interest. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors and was conducted using personal funding.

References

- [1] J. Marolla-Gajardo and I. Riquelme Plaza, "Teacher Education, Diversity, and the Prevention of Hate Speech: Ethical and Political Foundations for Inclusive Citizenship," *Societies*, vol. 15, no. 5, p. 139, 2025. doi: [10.3390/soc15050139](https://doi.org/10.3390/soc15050139).
- [2] B. Vidgen and L. Derczynski, "Directions in abusive language training data, a systematic review: Garbage in, garbage out.," *PLoS One*, vol. 15, no. 12, p. e0243300, 2020, doi: [10.1371/journal.pone.0243300](https://doi.org/10.1371/journal.pone.0243300).
- [3] P. Fortuna and S. Nunes, "A Survey on Automatic Detection of Hate Speech in Text," *ACM Comput. Surv.*, vol. 51, pp. 1–30, Jul. 2018, doi: [10.1145/3232676](https://doi.org/10.1145/3232676).
- [4] L. Floridi et al., "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations," *Minds Mach.*, vol. 28, no. 4, pp. 689–707, 2018, doi: [10.1007/s11023-018-9482-5](https://doi.org/10.1007/s11023-018-9482-5).
- [5] B. Mathew, P. Saha, C. Biemann, P. Goyal, and A. Mukherjee, "HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection," *Proc. AAAI Conf. Artif. Intell.*, vol. 35, pp. 14867–14875, May 2021, doi: [10.1609/aaai.v35i17.17745](https://doi.org/10.1609/aaai.v35i17.17745).
- [6] Ziqi Zhang and Lei Luo, "Hate speech detection: A solved problem? The challenging case of long tail on Twitter," *Semant. Web*, vol. 10, no. 5, pp. 925–945, Sep. 2019, doi: [10.3233/SW-18033](https://doi.org/10.3233/SW-18033).
- [7] M. Mozafari, R. Farahbakhsh, and N. Crespi, "A BERT-Based Transfer Learning Approach for Hate Speech Detection in Online Social Media BT - Complex Networks and Their Applications VIII," H. Cherifi, S. Gaito, J. F. Mendes, E. Moro, and L. M. Rocha, Eds., Cham: Springer International Publishing, 2020, pp. 928–940. doi: [10.1007/978-3-030-36687-2_77](https://doi.org/10.1007/978-3-030-36687-2_77).
- [8] B. Wilie et al., "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, Suzhou, China: Association for Computational Linguistics, 2020, pp. 843–857. doi: [10.18653/v1/2020.aacl-main.85](https://doi.org/10.18653/v1/2020.aacl-main.85).
- [9] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient Transformers: A Survey," *ACM Comput. Surv.*, vol. 55, no. 109, pp. 1–28, Sep. 2020, doi: [doi: 10.1145/3530811](https://doi.org/10.1145/3530811).
- [10] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *CoRR*, vol. abs/1910.01108, 2019, [Online]. Available: <http://arxiv.org/abs/1910.01108>
- [11] M.-L. Zhang and Z.-H. Zhou, "A Review on Multi-Label Learning Algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, 2014, doi: [10.1109/TKDE.2013.39](https://doi.org/10.1109/TKDE.2013.39).
- [12] F. M. Plaza-Del-Arco, M. D. González, L. Ureña-López, and M. Martín-Valdivia, "Integrating implicit and explicit linguistic phenomena via multi-task learning for offensive language detection," *Knowledge-Based Syst.*, vol. 258, p. 109965, Oct. 2022, doi: [10.1016/j.knsys.2022.109965](https://doi.org/10.1016/j.knsys.2022.109965).
- [13] 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).
- [14] T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Q. Liu and D. Schlangen, Eds., Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. doi: [10.18653/v1/2020.emnlp-demos.6](https://doi.org/10.18653/v1/2020.emnlp-demos.6).
- [15] S. Shaik and C. S P, "NLP and Machine Learning for Sentiment Analysis in COVID-19 Tweets: A Comparative Study," *EAI Endorsed Trans. Pervasive Heal. Technol.*, vol. 10 SE-Research article, doi: [10.4108/etpht.10.7051](https://doi.org/10.4108/etpht.10.7051).
- [16] D. W. Brata, A. Djunaidy, and D. O. Siahaan, "Comparison of Deep Learning Methods in Detecting Hate Speech in Indonesian Tweets," in *Proceedings of the 8th International Conference on Sustainable Information Engineering and Technology*, in SIET '23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 58–63. doi: [10.1145/3626641.3626925](https://doi.org/10.1145/3626641.3626925).
- [17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Association for Computational Linguistics*, Jun 2019,

- pp. 4171–4186. 2018. doi: [10.18653/v1/N19-1423](https://doi.org/10.18653/v1/N19-1423).
- [18] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, “IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP,” in *Proceedings of the 28th International Conference on Computational Linguistics*, D. Scott, N. Bel, and C. Zong, Eds., Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 757–770. doi: [10.48550/arXiv.2011.00677](https://doi.org/10.48550/arXiv.2011.00677).
- [19] F. Koto, J. H. Lau, and T. Baldwin, “IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W. Yih, Eds., Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 10660–10668. doi: [10.18653/v1/2021.emnlp-main.833](https://doi.org/10.18653/v1/2021.emnlp-main.833).
- [20] A. Paszke et al., “PyTorch: an imperative style, high-performance deep learning library,” in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA: Curran Associates Inc., 2019. doi: [10.48550/arXiv.1912.01703](https://doi.org/10.48550/arXiv.1912.01703).
- [21] T. Davidson, D. Warmusley, M. Macy, and I. Weber, “Automated Hate Speech Detection and the Problem of Offensive Language,” *Proc. Int. AAAI Conf. Web Soc. Media*, vol. 11, pp. 512–515, May 2017, doi: [10.1609/icwsm.v11i1.14955](https://doi.org/10.1609/icwsm.v11i1.14955).
- [22] Z. Waseem and D. Hovy, “Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter,” in *Proceedings of the {NAACL} Student Research Workshop*, J. Andreas, E. Choi, and A. Lazaridou, Eds., San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 88–93. doi: [10.18653/v1/N16-2013](https://doi.org/10.18653/v1/N16-2013).
- [23] A. Schmidt and M. Wiegand, “A Survey on Hate Speech Detection using Natural Language Processing,” in *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, L.-W. Ku and C.-T. Li, Eds., Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 1–10. doi: [10.18653/v1/w17-1101](https://doi.org/10.18653/v1/w17-1101).
- [24] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, “Deep Learning for Hate Speech Detection in Tweets,” in *Proceedings of the 26th International Conference on World Wide Web Companion*, in WWW ’17 Companion. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2017, pp. 759–760. doi: [10.1145/3041021.3054223](https://doi.org/10.1145/3041021.3054223).
- [25] M. Weber, C. Viehmann, M. Ziegele, and C. Scherer, “Online Hate Does Not Stay Online – How Implicit and Explicit Attitudes Mediate the Effect of Civil Negativity and Hate in User Comments on Prosocial Behavior,” *Comput. Human Behav.*, vol. 104, p. 106192, 2020, doi: <https://doi.org/10.1016/j.chb.2019.106192>.
- [26] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, “Abusive Language Detection in Online User Content,” in *Proceedings of the 25th International Conference on World Wide Web*, in WWW ’16. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2016, pp. 145–153. doi: [10.1145/2872427.2883062](https://doi.org/10.1145/2872427.2883062).
- [27] V. Basile et al., “SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter,” in *Proceedings of the 13th International Workshop on Semantic Evaluation*, J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, and S. M. Mohammad, Eds., Minneapolis, Minnesota, USA: Association for Computational Linguistics, Jun. 2019, pp. 54–63. doi: [10.18653/v1/S19-2007](https://doi.org/10.18653/v1/S19-2007).
- [28] A.-M. Founta et al., “Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior,” *Proc. Int. AAAI Conf. Web Soc. Media*, vol. 12, no. 1, Feb. 2018, doi: [10.48550/arXiv.1802.00393](https://doi.org/10.48550/arXiv.1802.00393).
- [29] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar, “Predicting the Type and Target of Offensive Posts in Social Media,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds., Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 1415–1420. doi: [10.18653/v1/N19-1144](https://doi.org/10.18653/v1/N19-1144).
- [30] S. Rosenthal, N. Farra, and P. Nakov, “SemEval-2017 Task 4: Sentiment Analysis in Twitter,” in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, S. Bethard, M. Carpuat, M. Apidianaki, S. M. Mohammad, D. Cer, and D. Jurgens, Eds., Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 502–518. doi: [10.18653/v1/S17-2088](https://doi.org/10.18653/v1/S17-2088).