Jurnal Minfo Polgan (JMP)
Volume 15, Nomor 2, Mei-Agustus 2026
DOI : https://doi.org/10.33395/jmp.v15i2.16581
e-ISSN : 2797-3298
p-ISSN : 2089-9424

Indonesian Hate Speech Detection Across Diverse Domains Using Parameter-Efficient Fine-Tuning with IndoBERT and LoRA

[1*]Fergie Joanda Kaunang, [2]Bhustomy Hakim, [3]Angelina Pramana Thenata
[1*,3]Department of Informatics, Faculty of Technology and Design, Bunda Mulia University, Jakarta, Indonesia
[2]Department of Information Systems, Faculty of Technology and Design, Bunda Mulia University, Jakarta, Indonesia
*Corresponding Author:: fkaunang@bundamulia.ac.id

Submitted: June 28, 2026 | **Accepted:** July 20, 2026 | **Published:** Aug 10, 2026

ABSTRACT

The rapid proliferation of digital connectivity in Indonesia has catalyzed an unprecedented surge in harmful online content, necessitating robust automated systems for hate speech detection that can generalize across diverse digital platforms. Traditional models often struggle with domain shift and the linguistic complexities of Indonesian social media discourse, including informal slang and code-mixing. This research proposes a multi-domain detection framework leveraging the IndoBERT-base-p1 architecture integrated with Low-Rank Adaptation (LoRA), a parameter-efficient fine-tuning (PEFT) strategy. The study utilizes a multi-source corpus from Instagram, Twitter, and news portals, employing back-translation to augment scarce Instagram data and stratified downsampling to ensure domain equilibrium. By training only 1–2% of the total 110 million parameters, specifically targeting the query and value attention modules, the model achieves significant computational savings with a training loss of 0.345. Experimental results demonstrate high robustness, with the framework attaining F1-scores of 0.83 for both Instagram and Twitter, and 0.81 for news portals, while maintaining accuracies between 0.81 and 0.85. Qualitative validation through word cloud analysis further confirms the model's ability to distinguish between aggressive sociopolitical triggers and neutral functional discourse. This study contributes a scalable and resource-efficient solution for real-time content moderation, proving effective across both formal journalistic Indonesian and informal digital dialects. The findings indicate that IndoBERT+LoRA provides a promising and resource-efficient approach for multi-domain Indonesian hate and abusive speech classification, while stricter leave-one-domain-out evaluation remains an important direction for future work.

Keywords: Back Translation, IndoBERT, Indonesian Hate Speech, Low-Rank Adaptation (LoRA), Multi-domain Classification

ABSTRAK

Seiring dengan pesatnya pertumbuhan konektivitas digital di Indonesia, penyebaran konten berbahaya di ruang daring juga semakin meningkat. Oleh karena itu, diperlukan sistem deteksi ujaran kebencian otomatis yang tidak hanya akurat, tetapi juga mampu beradaptasi dengan karakteristik bahasa pada berbagai platform digital. Model tradisional seringkali kesulitan menghadapi pergeseran domain (domain shift) dan kompleksitas linguistik dalam media sosial di Indonesia, termasuk penggunaan bahasa gaul (alay) dan pencampuran bahasa (code-mixing). Dalam penelitian ini, istilah multi-domain merujuk pada penggunaan beberapa sumber teks berbahasa Indonesia, yaitu Instagram, Twitter, dan portal berita, yang digabungkan dalam satu kerangka klasifikasi dan dievaluasi secara global maupun per domain. Penelitian ini mengusulkan kerangka kerja deteksi multi-domain menggunakan arsitektur IndoBERT yang diintegrasikan dengan Low-Rank Adaptation (LoRA), sebuah strategi fine-tuning yang efisien secara parameter. Studi ini memanfaatkan korpus multi-sumber dari Instagram, Twitter, dan portal berita, serta menerapkan teknik back-translation untuk menambah data Instagram yang terbatas dan stratified downsampling untuk memastikan keseimbangan antar domain. Dengan melatih hanya 1–2% dari total 110 juta parameter, khususnya pada modul atensi query dan value, model ini mencapai penghematan komputasi yang signifikan dengan training loss sebesar 0,345. Hasil eksperimen menunjukkan ketahanan yang tinggi, di mana kerangka kerja ini meraih

skor F1 sebesar 0,83 untuk Instagram dan Twitter, serta 0,81 untuk portal berita, dengan tingkat akurasi antara 0,81 hingga 0,85. Hasil penelitian menunjukkan bahwa IndoBERT+LoRA merupakan pendekatan yang menjanjikan dan efisien untuk klasifikasi ujaran kebencian dan abusif bahasa Indonesia pada data multi-domain, sedangkan evaluasi leave-one-domain-out yang lebih ketat dapat menjadi arah penelitian lanjutan.

***Kata Kunci**: Back Translation, IndoBERT, Klasifikasi Multi Domain, Low-Rank Adaptation (LoRA), Ujaran Kebencian Indonesia*

INTRODUCTION

The digital ecosystem in Indonesia has undergone a profound transformation, characterized by an unprecedented surge in internet penetration and social media engagement. By 2023, internet penetration had reached 77.2% of the population, a figure that continues to climb as digital infrastructure expands into the more remote regions of the archipelago (Sonni, 2025). As of 2025, the number of internet users in Indonesia is estimated at approximately 220 million, making it one of the most active and volatile digital markets globally (Kholik & Sulastri, 2025). This rapid proliferation of connectivity, while fostering democratic engagement and economic opportunity, has simultaneously created a fertile ground for the dissemination of harmful content. The spread of hoaxes, hate speech, and abusive language has emerged as a critical threat to social cohesion and national stability, particularly in the context of high-stakes political events such as the 2024 General Elections (Khuan & Wahyudi, 2024).

Automated hate speech detection is no longer merely a theoretical pursuit in the field of Natural Language Processing (NLP). It has become a fundamental necessity for maintaining a healthy digital ecosystem and ensuring compliance with Indonesia's evolving legal frameworks (Sipayung, 2024; Zikrina & Fitriyani, 2025). The Electronic Information and Transactions (ITE) Law, specifically Law No. 1 of 2024, represents a pivotal shift in the government's approach to digital governance. This law serves as a tool of social engineering designed to shape digital communication behavior through a combination of punitive measures and preventative mandates (Ibrohim & Budi, 2019). During the 2024 election period alone, authorities identified over 103,000 instances of negative content, underscoring the massive scale of the challenge (Kholik & Sulastri, 2025). However, the detection of such content is plagued by the complexities of the Indonesian language, which is characterized by a high degree of linguistic diversity, widespread code-mixing with regional dialects like Javanese and Sundanese, and a dynamic layer of informal slang known as "alay". For example, a user might replace letters with numbers or use metaphors to bypass keyword-based filters (Saputra & Sibaroni, 2025). Traditional models that rely on rigid word-level embeddings or TF-IDF often fail to recognize these variations (Putra & Nurjanah, 2020). Advanced preprocessing pipelines are now required to perform text normalization, converting "alay" words into formal equivalents before the classification process (Hanafi & Sriyanto, 2025; Winson & Hiskiawan, 2026).

The development of Indonesian hate speech detection has mirrored global trends in NLP, transitioning from classical machine learning to sophisticated deep learning and transformer-based architectures. Initial research focused on Support Vector Machines (SVM), Naive Bayes (NB), and Random Forest Decision Trees (RFDT) (Ibrohim & Budi, 2019). These models relied heavily on manual feature extraction, such as word unigrams, character quad-grams, and lexicon-based features. While efficient, these methods achieved limited accuracy, often hovering around 66% to 77% on complex multi-label datasets (Ibrohim & Budi, 2019). One notable study on long-form documents from Facebook found that combining SVM with TF-IDF and char quad-grams yielded an F1-score of 85%, suggesting that classical methods still have utility for specific text types (Aulia & Budi, 2019).

The shift toward deep learning introduced transformer-based architectures. Bidirectional Encoder Representations from Transformers (BERT) has set a new standard for Indonesian text classification (Winson & Hiskiawan, 2026). Models like IndoBERT and IndoBERTweet are pre-trained on massive Indonesian corpora, allowing them to capture deep semantic and syntactic structures (Hanafi & Sriyanto, 2025; Yuswira & Ginting, 2026). Previous study combines the transformer layers with a 64-unit BiLSTM layer, researchers achieved an accuracy of 93.7%. The BiLSTM layer allows the model to better capture sequential patterns that might be missed by the self-attention mechanism alone (Kusuma & Chowanda, 2023). Another study introducing Graph Neural Networks (GNN) for feature extraction achieved 92.9% accuracy for abusive content, outperforming standard RNNs. GNNs are particularly effective at

capturing the complex relational nuances between words in a graph structure (Zikrina & Fitriyani, 2025). For shorter texts, such as brief tweets, BERT combined with CNN layers which excel at local feature extraction. The combined model reached an accuracy of 79.8%, proving more effective than BiLSTM for limited word sets (Saputra & Sibaroni, 2025).

A persistent problem in hate speech detection is the lack of generalization across different digital platforms. Most existing models are trained and tested on Twitter data, yet linguistic characteristics vary significantly across Instagram, Facebook, and news portals. Platforms like Instagram are characterized by a high volume of visual content where hate speech is predominantly located in the comments. About 60% of hate speech on Instagram is found in comments, often triggered by personal photos or videos (Putra & Nurjanah, 2020). In contrast, YouTube serves as a central arena for political issue diffusion, where mainstream media narratives are reinterpreted and contested through parasitic news and ideological framing (Amalia, Harani, & Prianto, 2024; Fajar Maulana, Rando, Hastuti, Zaitullah, & Zarifah Ferizka, 2025). To address these challenges, this study proposes a multi-domain Indonesian hate and abusive speech detection framework leveraging IndoBERT combined with Low-Rank Adaptation (LoRA). LoRA represents a Parameter-Efficient Fine-Tuning (PEFT) strategy that optimizes only a small fraction of the model's weights by injecting trainable low-rank matrices into the Transformer layers (Nwaiwu, 2025). Rather than claiming strict zero-shot cross-domain transfer, this study focuses on evaluating whether a single parameter-efficient model can maintain stable performance across heterogeneous Indonesian text domains when trained on an integrated multi-domain corpus. This study aims to evaluate the effectiveness of PEFT in maintaining high classification accuracy while significantly reducing computational overhead. By evaluating the model both globally and per domain, this research seeks to provide empirical evidence on the consistency and efficiency of IndoBERT+LoRA for Indonesian hate and abusive speech classification across diverse digital text sources.

RESEARCH METHOD

This research uses a deep learning-based computational approach with a focus on parameter efficiency for multi-domain Indonesian hate and abusive speech classification. As shown in Figure 1, the workflow of this study is divided into four main stages: data acquisition, data preprocessing, architectural modeling with LoRA integration, and evaluation procedures.

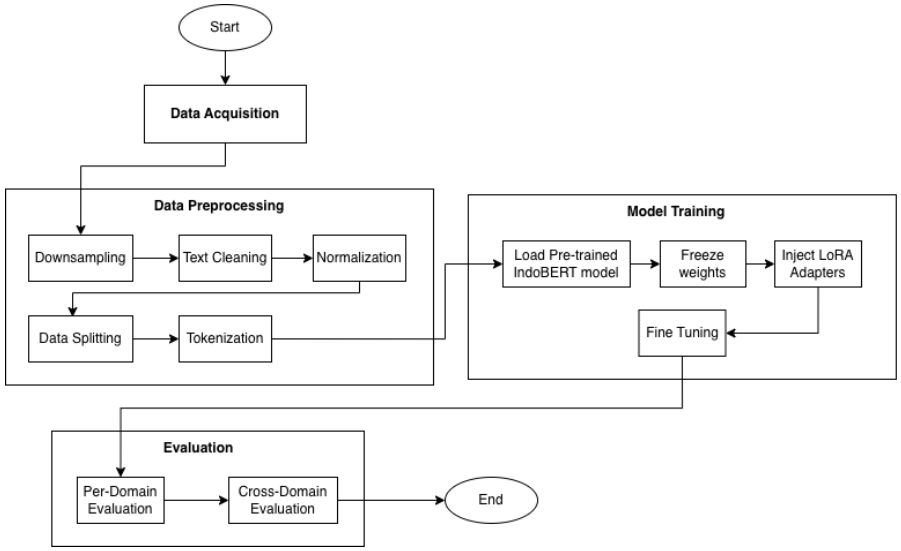

Figure 1 Research Flow

Data Acquisition and Augmentation

The study utilizes a multi-source corpus comprising datasets from Twitter (Ibrohim & Budi, 2023; Tonneau et al., 2024), Instagram (Pratiwi, Budi, & Jiwanggi, 2019; Tonneau et al., 2024), and News Portals (Firdaus & Rizki, 2024). These three sources were treated as distinct domains because they represent different communication settings. Twitter data reflects short and informal public discourse, Instagram data represents comment-based interaction often associated with visual content, while news portal data contains longer and relatively more formal

responses to public issues. Therefore, the dataset design supports multi-domain evaluation by allowing model performance to be analyzed both globally and separately for each domain. To address the challenge of limited data availability particularly in the Instagram domain, Back Translation was employed as a data augmentation technique.

Table 1 Dataset distribution before train-validation-test splitting

Domain	Hate/Abusive	Non-Hate/Non-Abusive	Total
Instagram after augmentation	572	572	**1,144**
Twitter	7,309	5,860	**13,169**
News Portal	5,739	9,617	**15,356**
Total	**13,620**	**16,049**	**29,669**

The text was translated into English and subsequently back into Indonesian to generate synthetic samples, thereby increasing the diversity of the training set while maintaining semantic consistency. To reduce the risk of semantic leakage, augmented instances should be kept within the same data partition as their original counterparts. In the final experimental setting, augmentation was applied only to the training portion of the Instagram data, while validation and test sets were kept in their original form. Following augmentation, stratified splitting and domain-level downsampling were applied to reduce class and domain imbalance, particularly to prevent the larger Twitter and News Portal datasets from dominating the training process.

Data Preprocessing

The preprocessing stage consisted of several steps to ensure data quality and model stability. First, all text was converted to lowercase, and standard noise reduction was performed using regular expressions (Heksaputra, Gernowo, & Isnanto, 2024), including the removal of URLs, user mentions (@username), hashtags (#), non-alphanumeric characters such as emoticons and punctuation, and numerical digits. Second, informal Indonesian expressions and slang were normalized using a specialized Indonesian colloquial lexicon to reduce lexical variation commonly found in social media text. After cleaning and normalization, empty records were removed. The dataset was then split into training, validation, and test sets using stratified sampling based on label-domain combinations, ensuring that each domain and class remained represented in all partitions. A total of 20% of the data was reserved as a hold-out test set, while 10% of the remaining data was used as validation data. Downsampling was applied only to the training set by limiting the majority domains, Twitter and News Portal, to approximately 3,000 samples each, while retaining all available Instagram samples. Finally, the text was tokenized using the BertTokenizer from indobenchmark/indobert-base-p1, with a maximum sequence length of 128 tokens and padding/truncation applied to produce uniform input tensors.

Hyperparameter Optimization and Model Training

To adapt the pre-trained IndoBERT model to the hate speech classification task, we utilized Low-Rank Adaptation (LoRA), a strategy that offers significant computational savings (Hidayatullah, 2024). This research employs the indobert-base-p1 model as the foundational architecture. Instead of retraining all 110 million+ parameters, the base model weights were frozen, and Low-Rank Adaptation (LoRA) matrices were injected. This approach targets the query and value modules within the attention layers, training only approximately 1–2% of the total parameters. A grid search was conducted to identify the optimal configuration (Simanjuntak et al., 2024). The optimization process was executed in two systematic stages. First, a grid search was performed to evaluate various combinations of LoRA-specific parameters, specifically targeting rank values (r) of {4,8,16,32} and scaling factors (α) of {16,32,64}. The F1-score was utilized as the primary metric to determine the best-performing configuration, ensuring the model could effectively handle class imbalances in hate speech detection.

Second, after identifying the optimal LoRA rank and scaling factor, the training dynamics were further refined by tuning the learning rate and the number of epochs. The learning rates tested were {1e-4, 2e-4, 5e-4}, while the epoch counts were {3, 5, 10, 15, 20}. To prevent

overfitting and ensure better generalization, an EarlyStoppingCallback with a patience of 3 epochs was implemented by monitoring the validation F1-score. The final configuration used LoRA with r = 4 and α = 32, targeting the query and value modules of the IndoBERT attention layers. The model was trained using the AdamW optimizer, a batch size of 8, and mixed precision training (FP16) on an NVIDIA T4 GPU. The final parameter configuration can be seen in Table 2.

Table 2 Final training and LoRA configuration

Parameter	Value
Base model	indobenchmark/indobert-base-p1
Task	Binary sequence classification
Number of labels	2
Max length	128
Optimizer	AdamW
Batch size	8
Epochs	5
LoRA rank (r)	4
LoRA alpha (α)	32
LoRA dropout	0.1
Target modules	query, value
Mixed precision	FP16
GPU	NVIDIA T4

Evaluation

The performance of the fine-tuned model was evaluated through two complementary perspectives: global multi-domain evaluation and per-domain evaluation. In the global evaluation, the model was tested on the integrated hold-out test set containing samples from Instagram, Twitter, and News Portal domains. This evaluation measures the overall classification performance of the model on heterogeneous Indonesian text sources. In the per-domain evaluation, the same hold-out test set was filtered by domain to analyze whether the model maintained consistent performance across different digital environments. Therefore, the evaluation in this study is positioned as multi-domain evaluation rather than strict leave-one-domain-out cross-domain transfer. Quantitative performance was measured using Accuracy, Precision, Recall, and F1-Score, supported by confusion matrices for detailed error analysis. Word cloud visualization was used as exploratory lexical analysis to compare frequent terms in hate/abusive and non-hate/non-abusive text.

RESULTS AND DISCUSSIONS

The initial phase of this research focused on rectifying the severe data imbalance across Indonesian digital platforms. As shown in Figure 2, the raw data was heavily skewed toward Twitter and News Portal domains, leaving the Instagram domain significantly underrepresented.

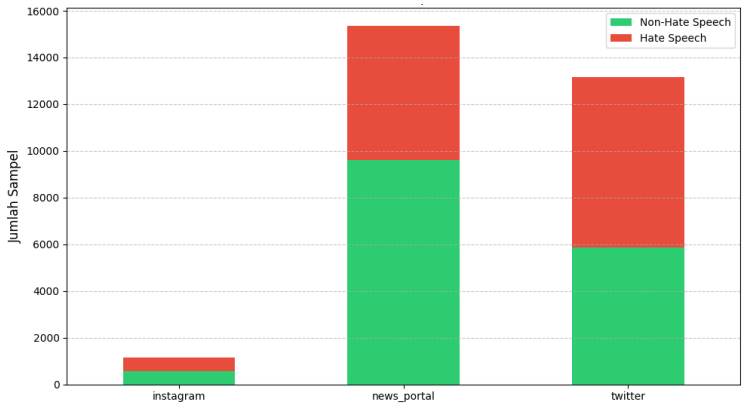

Figure 2 Distribution of Hate Speech Labels Per Domain

To resolve this, a Back Translation augmentation was applied to the Instagram dataset, effectively doubling its size from 572 to 1,144 samples. By translating text from Indonesian to English and back, we introduced syntactic variety while preserving semantic offensive intent. Following augmentation and a stratified downsampling strategy, which capped majority domains at approximately 3,000 samples, the final dataset distribution reached a state of equilibrium suitable for robust transformer training. The adoption of Low-Rank Adaptation (LoRA) proved to be highly efficient for the Indonesian linguistic context. By injecting low-rank matrices into the query and value attention modules of the IndoBERT-base-p1 model, we updated only approximately 1–2% of the total 110 million parameters. The model demonstrated rapid convergence, achieving a training loss of 0.345 within 5 epochs. This high level of efficiency indicates that hate speech classification patterns in Indonesian can be captured within a low-rank subspace, significantly reducing the computational and storage costs compared to full-model fine-tuning.

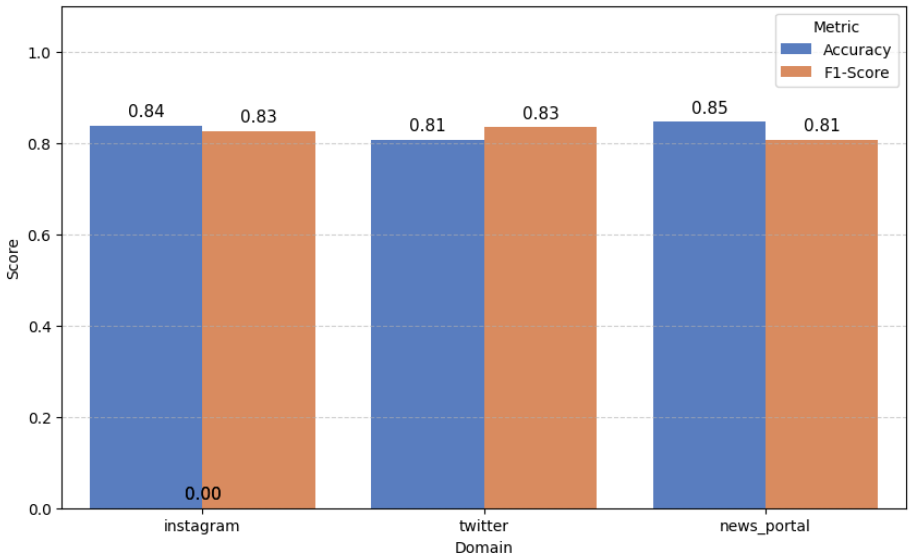

Figure 3 Model Performance (Accuracy & F1-Score) Across Domains

The model's performance was evaluated through a multi-domain lens to ensure it could handle the stylistic shift between formal news and informal social media discourse. As visualized in Figure 3, the results were consistently high. In the Instagram domain, the classification model demonstrated robust performance, achieving an accuracy of 0.84 and an F1-Score of 0.83. For the Twitter dataset, the model maintained consistent reliability with an accuracy of 0.81 and an F1-Score of 0.83. Furthermore, the model yielded its highest accuracy of 0.85 in the News Portal domain, accompanied by an F1-Score of 0.81. These results suggest that the augmented Instagram data successfully provided the model with enough representative patterns to match the performance of naturally larger datasets. It should be noted that these scores represent performance under a multi-domain hold-out evaluation setting, where all three domains were represented during training and testing. Therefore, the results demonstrate performance consistency across heterogeneous domains rather than strict zero-shot cross-domain generalization.

To understand the model's reliability, we analyzed the Confusion Matrices for both individual domains and the global test set. In the Global Confusion Matrix shown in Figure 4, the model correctly identified 2,558 instances of Non-Hate Speech and 2,356 instances of Hate Speech.

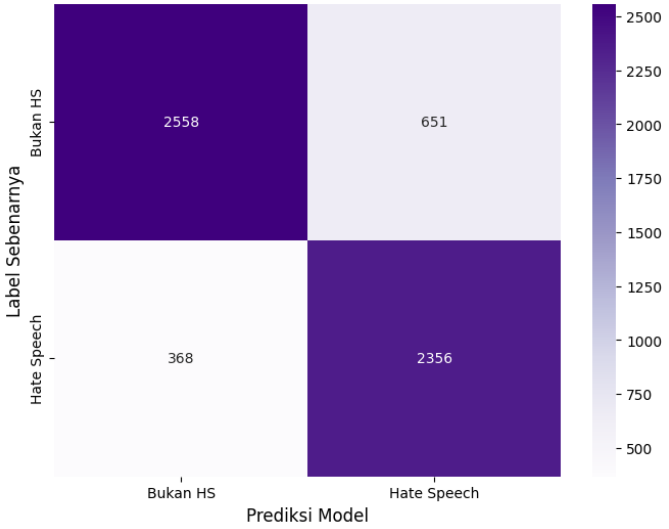

Figure 4 Global Confusion Matrix

The matrices reveal that while the model is highly sensitive to offensive content, minimizing False Negatives, there is a slight tendency toward False Positives (651 instances globally). The relatively higher number of false positives indicates that the model tends to be conservative when detecting potentially offensive content. This tendency may occur because several non-hate comments contain harsh expressions, political keywords, or emotionally charged words that resemble hate/abusive speech patterns. In practical moderation settings, this behavior should be interpreted carefully because excessive false positives may lead to the removal of legitimate criticism or neutral discussion. Therefore, future deployment should include human review for borderline cases, especially in politically sensitive domains.

The final model configuration (r=4, α=32) was the result of extensive hyperparameter tuning. We further optimized the training by testing various learning rates and epoch counts, implementing an EarlyStoppingCallback with a patience of 3 epochs. This ensured that the model stopped training at the point of maximum validation F1-score, effectively preventing overfitting and ensuring the reported metrics reflect the most generalized version of the model. To complement the quantitative evaluation, a frequency-based exploratory lexical analysis was conducted using Word Cloud visualizations (Figures 5 and 6). This analysis does not directly explain the internal decision-making process of the model, but it provides an overview of dominant lexical patterns in hate/abusive and non-hate/non-abusive samples.

Figure 5 Word Cloud: Top 100 Terms in Hate Speech

The top 100 terms in Hate Speech shown in Figure 5, reveals a high density of explicit derogatory lexical items commonly used in Indonesian social media, such as "tolol", "goblok", and "anjing". Notably, the visualization also highlights the model's sensitivity to culturally situated sociopolitical slurs, including terms like "cebong", "kadrun", and references to political figures. The prominence of these keywords confirms that the IndoBERT+LoRA architecture successfully learned to prioritize high-impact offensive markers that characterize Indonesian hate speech across various digital platforms. In contrast, the top 100 terms in normal text (see Figure 6) is dominated by functional and neutral terms. The prevalence of words such as "dan"

(and), "yang" (which/that), "untuk" (for), and "ini" (this) reflects the standard grammatical structure of the Indonesian language. The absence of aggressive triggers in this cloud suggests that the model's evaluation of neutral text is rooted in general conversational flow rather than specific offensive patterns.

Figure 6 Word Cloud: Top 100 Terms in Normal Text

A notable observation in both word clouds is the presence of common functional words and placeholders like "user". The inclusion of these elements is a deliberate methodological choice rather than a deficiency in preprocessing. Unlike traditional frequency-based models, the IndoBERT-base-p1 architecture relies on bidirectional attention mechanisms to understand the relationship between words. Retaining functional words or stopwords is critical for maintaining the syntactic structural integrity of the sentences, allowing the model to capture nuances in meaning that would be lost if the text were reduced to isolated keywords. The recurrent term "user" results from the systematic cleaning of mentions (@username) during the preprocessing stage to ensure data anonymity and reduce the high-cardinality noise associated with unique usernames. The distinction between aggressive markers in Figure 5 and neutral functional words in Figure 6 supports the quantitative findings by showing that hate/abusive samples contain more explicit derogatory and sociopolitical terms than non-hate/non-abusive samples. However, because word clouds are frequency-based, they should be interpreted as exploratory evidence rather than a direct explanation of model attention or decision-making. This balance is instrumental in achieving the reported Accuracy of 0.81–0.85 and F1-Scores of 0.81–0.83 across multi-domain datasets, proving that the model is robust enough to navigate the complexities of both formal and informal Indonesian language.

CONCLUSION

This research demonstrates the potential of integrating Parameter-Efficient Fine-Tuning (PEFT) via Low-Rank Adaptation (LoRA) with the IndoBERT-base-p1 architecture for multi-domain Indonesian hate and abusive speech classification. By targeting the query and value modules of the attention layers, the model achieved stable performance while training only a small fraction of the total parameters, thereby reducing computational overhead. The use of back-translation for the Instagram domain, combined with stratified splitting, normalization, and training-set downsampling, helped mitigate data scarcity and domain imbalance in the multi-source corpus. Quantitative evaluation showed consistent performance across the investigated domains, with F1-Scores of 0.83 on Instagram, 0.83 on Twitter, and 0.81 on News Portal data. These results indicate that IndoBERT+LoRA can serve as an efficient approach for multi-domain Indonesian hate and abusive speech classification. Nevertheless, the evaluation setting in this study should be interpreted as multi-domain hold-out evaluation rather than strict cross-domain transfer, since all domains were represented in the training and test partitions. Therefore, further evaluation using leave-one-domain-out scenarios is recommended to assess true domain generalization. Future work should also conduct stricter cross-domain experiments, such as leave-one-domain-out evaluation, where one domain is completely excluded from training and used only for testing. This setting would provide stronger evidence of whether the model can generalize to unseen digital platforms.

REFERENCES

Amalia, F. R., Harani, N. H., & Prianto, C. (2024). Sentiment Analysis of Hate Speech Against Presidential Candidates of the Republic of Indonesia in the 2024 Election Using BERT. Jurnal Sistem Cerdas, 7(3), 339–355.

Aulia, N., & Budi, I. (2019). Hate speech detection on Indonesian long text documents using machine learning approach. Proceedings of the 2019 5th International Conference on Computing and Artificial Intelligence, 164–169.

Fajar Maulana, H., Rando, R., Hastuti, H., Zaitullah, L. O. M., & Zarifah Ferizka, Z. (2025). Formulation of a model for the dissemination of government policy issues in online media and YouTube in Indonesia. Frontiers in Communication, 10, 1710197.

Firdaus, M., & Rizki, P. N. M. (2024). BIJAKAWEB: Platform Berbasis Web Untuk Deteksi Hate Speech Pada Komentar Berita Bahasa Indonesia. Jurnal Teknologi Informasi Dan Ilmu Komputer, 11(4), 939–948.

Hanafi, M., & Sriyanto, A. (2025). Indonesian News Classification on Detik. com using LSTM-RNN, TF-IDF, and BERT Methods: Comparative Analysis of Performance. 2025 International Conference on Artificial Intelligence's Future Implementations (ICAIFI), 199–204. IEEE.

Heksaputra, D., Gernowo, R., & Isnanto, R. R. (2024). Model for Bahasa text cleaning based regular expression non-deterministic finite automata approach. 2024 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS), 171–176. IEEE.

Hidayatullah, A. F. (2024). Parameter efficient fine-tuning using low-rank adaptation for emotion classification in indonesian texts. 2024 9th International Conference on Information Technology and Digital Applications (ICITDA), 1–6. IEEE.

Ibrohim, M. O., & Budi, I. (2019). Multi-label hate speech and abusive language detection in Indonesian Twitter. Proceedings of the Third Workshop on Abusive Language Online, 46–57.

Ibrohim, M. O., & Budi, I. (2023). Hate speech and abusive language detection in Indonesian social media: Progress and challenges. Heliyon, 9(8).

Kholik, M. A., & Sulastri, D. (2025). Social Engineering Through Criminal Law: The Effectiveness of the ITE Law in Shaping Digital Communication in Indonesia. Ius Poenale, 6(2), 101–112.

Khuan, H., & Wahyudi, F. S. (2024). Juridical Study of Digital Campaign Regulations and Election Violations in the 2024 Elections in Indonesia_ Analysis of the Role of the ITE Law in Handling Hoaxes and Hate Speech. West Science Law and Human Rights, 3(1).

Kusuma, J. F., & Chowanda, A. (2023). Indonesian hate speech detection using IndoBERTweet and BiLSTM on Twitter. JOIV: International Journal on Informatics Visualization, 7(3), 773–780.

Nwaiwu, S. (2025). Parameter-efficient fine-tuning for low-resource text classification: a comparative study of LoRA, IA3, and ReFT. Frontiers in Big Data, 8, 1677331.

Pratiwi, N. I., Budi, I., & Jiwanggi, M. A. (2019). Hate speech identification using the hate codes for Indonesian tweets. Proceedings of the 2019 2nd International Conference on Data Science and Information Technology, 128–133.

Putra, I. G. M., & Nurjanah, D. (2020). Hate speech detection in Indonesian language Instagram. 2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS), 413–420. IEEE.

Saputra, R. A., & Sibaroni, Y. (2025). Multilabel Hate Speech Classification in Indonesian Political Discourse on X using Combined Deep Learning Models with Considering Sentence Length. Jurnal Ilmu Komputer Dan Informasi, 18(1), 113–125.

Simanjuntak, A., Lumbantoruan, R., Sianipar, K., Gultom, R., Simaremare, M., Situmeang, S., & Panggabean, E. (2024). Research and analysis of indobert hyperparameter tuning in fake news detection. Jurnal Nasional Teknik Elektro Dan Teknologi Informasi, 13(1), 60–67.

Sipayung, E. M. (2024). Sentiment on Public Trust Using the NLP Rule Based Method. JUSTIN (Jurnal Sistem Dan Teknologi Informasi), 12(1), 175–182.

Sonni, A. F. (2025). AI-based disinformation and hate speech amplification: analysis of Indonesia's digital media ecosystem. Frontiers in Communication, 10, 1603534.

Tonneau, M., Liu, D., Fraiberger, S., Schroeder, R., Hale, S. A., & Röttger, P. (2024). From languages to geographies: Towards evaluating cultural bias in hate speech datasets. Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024), 283–311.

Winson, W., & Hiskiawan, P. (2026). An Applied Data Science Approach for Detecting Depression Symptoms in Indonesian Social Media Text Using Transformer Models. Journal of Applied Informatics and Computing, 10(3), 2115–2127.

Yuswira, J. D., & Ginting, J. A. (2026). ANALISIS SENTIMEN DAN SOCIAL NETWORK ANALYSIS PADA ISU MAKANAN BERGIZI GRATIS DI X/TWITTER MENGGUNAKAN INDOBERT. JATI (Jurnal Mahasiswa Teknik Informatika), 10(1), 1259–1265.

Zikrina, S., & Fitriyani. (2025). Advancing Hate Speech Detection in Indonesian Language Using Graph Neural Networks and TF-IDF. Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi), 9, 137–145. https://doi.org/10.29207/resti.v9i1.6179