Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control
Journal homepage: http://kinetik.umm.ac.id
ISSN: 2503-2267
Vol. 11, No. 3, August, Pp. 503-512

Hate speech analysis of YouTube comments on the 2024 Indonesian presidential debate using IndoBERT

Agus Sasmito Aribowo *[1], Yuli Fauziah[1], Yusna Bantulu[2], Shoffan Saifullah[3], Azfa Mutiara Ahmad Fabulo[4]
Department of Informatics, Universitas Pembangunan Nasional "Veteran" Yogyakarta, Indonesia[1]
Department of Accounting, Universitas Pembangunan Nasional "Veteran" Yogyakarta, Indonesia[2]
Faculty of Computer Science, AGH University of Krakow, Poland[3]
Department of Accounting, Mercu Buana University Yogyakarta, Indonesia[4]

Article Info
Keywords:
Hate Speech Detection, IndoBERT, 2024 Indonesian Presidential Election, Semi-Supervised Learning

Article history:
Received: December 02, 2025
Accepted: April 27, 2026
Published: August 01, 2026

Cite:
A. S. Aribowo, Yuli Fauziah, Y. Bantulu, Shoffan Saifullah, and A. M. A. Fubalo, "Hate Speech Analysis of YouTube Comments on the 2024 Indonesian Presidential Debate Using IndoBERT", *KINETIK*, vol. 11, no. 3, Aug. 2026.
https://doi.org/10.22219/kinetik.v11i3.2604

*Corresponding author.
Agus Sasmito Aribowo
E-mail address:
sasmito.skom@upnyk.ac.id

Abstract
The rapid digitization of political campaigns has intensified the spread of hate speech on social media, threatening democratic discourse and social cohesion. In Indonesia, YouTube comments on the 2024 presidential election debates have emerged as a critical yet underexplored source of polarizing content. However, existing detection systems struggle with Indonesian-specific linguistic features, code-mixing, and implicit political sarcasm, while high annotation costs and class imbalance further limit the scalability of supervised approaches. To address these challenges, this study introduces a large-scale dataset of 38,742 YouTube comments collected from the five official debate stages and labeled using a cost-effective semi-supervised framework (20% expert-annotated, 80% pseudo-labeled). We systematically evaluate four classification models —IndoBERT, mBERT, SVM, and Random Forest—under identical experimental conditions using evaluation metrics optimized for imbalanced data. Experimental results demonstrate that IndoBERT consistently outperforms all baseline models, achieving an average accuracy of 89.7% and a macro F1-score of 0.89 across all debate stages. Notably, IndoBERT maintains high recall for the minority hate speech class (0.81–0.90), confirming its superior ability to capture localized political rhetoric and contextual nuances that multilingual and classical models frequently miss. This study contributes a publicly available Indonesian political hate speech dataset, validates a scalable semi-supervised annotation pipeline, and provides empirical evidence that domain-specific Transformer models are essential for reliable content moderation in politically charged, low-resource environments.

1. Introduction

Social media has fundamentally transformed political communication worldwide, serving as a primary campaign tool for shaping public opinion and mobilizing support [1]–[4]. However, this digital shift has simultaneously amplified the proliferation of hate speech, threatening democratic discourse and social cohesion [5]. Globally, automated hate speech detection has evolved from rudimentary rule-based keyword filtering to sophisticated machine learning approaches. Recent advancements have been dominated by Transformer-based architectures, particularly BERT and its variants [6], [7], which leverage self-attention mechanisms to capture long-range semantic dependencies and enable a nuanced understanding of context, sarcasm, and implicit aggression. In the Indonesian context, IndoBERT—a model pre-trained on large-scale Indonesian corpora—has consistently demonstrated superior performance over multilingual counterparts (e.g., mBERT) across various NLP tasks, including sentiment analysis [8], [9] aspect-based opinion mining [10], and domain-specific text classification [11], [12]. This technological progress coincides with a critical period in Indonesia's democratic landscape: the 2024 presidential election. During this period, YouTube emerged as a primary platform for broadcasting official debates, becoming a focal point for public discourse and a conduit for polarizing content, including identity-based attacks, sarcastic rhetoric, and SARA-related provocations.

Despite these advancements, three critical challenges persist in detecting hate speech within Indonesian political discourse. First, most existing studies rely on multilingual models like mBERT, which, although broadly applicable, frequently underperform in capturing Indonesian-specific linguistic nuances, including code-mixing, local political slang (e.g., "kadrun," "cebong," "buzzer"), and culturally embedded sarcasm [7], [9]. The lack of domain-specific pre-training limits their ability to distinguish legitimate political criticism from genuine hate speech. Second, high-quality labeled data remain scarce and prohibitively costly to produce. Manual annotation requires linguistic expertise, rigorous inter-annotator agreement protocols, and substantial time investment [13]–[16], creating a bottleneck for low-resource languages and rapidly evolving political contexts where rapid model deployment is essential. Third, prior research on Indonesian political text has predominantly focused on three-class sentiment analysis (positive/neutral/negative) rather

than explicit binary hate speech classification [8], [9]. Moreover, few studies address the severe class imbalance inherent in real-world moderation tasks or provide comparative benchmarking against classical machine learning baselines under semi-supervised learning settings [17]. Collectively, these gaps hinder the development of robust, scalable, and context-aware hate speech detection systems for Indonesian digital platforms during critical democratic periods [18].

To address these challenges, this study introduces a novel, integrated framework that advances both methodological rigor and practical applicability in Indonesian hate speech detection. First, we present a large-scale, publicly available dataset comprising 38,742 YouTube comments collected from official videos covering the five stages of the 2024 Indonesian presidential debate. The dataset is annotated using a cost-effective semi-supervised pipeline: 20% is expert-labeled by three independent assessors (Fleiss' Kappa = 0.78), while the remaining 80% is pseudo-labeled using a self-training framework [18], [19]. This strategy significantly reduces annotation costs while preserving label quality and model generalizability. Second, we conduct a comprehensive and fair comparative evaluation of four classification models—IndoBERT, mBERT, SVM, and Random Forest—under identical experimental conditions, using evaluation metrics optimized for imbalanced data, including accuracy, precision, recall, and macro F1-score. Third, we empirically validate a semi-supervised learning approach that achieves competitive detection performance using only 20% labeled data, offering a scalable pathway for deploying moderation systems in resource-constrained environments. The novelty of this work lies in combining a domain-specific Transformer model with a scalable semi-supervised annotation pipeline, explicitly addressing class imbalance and localized political rhetoric to deliver a high-performance, real-world-ready solution for content moderation. The remainder of this paper is organized as follows. Section 2 describes the research methodology, including data collection, preprocessing, labeling, and model configuration. Section 3 presents the experimental results and discussion, including a comparative analysis across debate stages and classification models. Finally, Section 4 concludes the paper by discussing its implications for policy, platform governance, and future research.

2. Research Method

2.1 Data Collection

The dataset was collected from comments posted by viewers on official YouTube videos of the presidential and vice-presidential candidates uploaded during January–February 2024. The data crawling process was divided into five stages, corresponding to the five presidential debate stages. For each debate stage, comments were extracted from several videos related to the debate. Selenium functions implemented in Python were used to extract comments from the debate videos. The extracted comments were then manually cleaned. Only comments related to the debate videos were retained, while comments containing advertising, promotional content, or content unrelated to the videos were removed. The results of the initial data collection and cleaning process are presented in Table 1.

Table 1. Data Collection and Cleaning Results by Debate Stage

Stage	Date	Number of Instances
1	14 January 2024	8,215
2	21 January 2024	7,632
3	28 January 2024	7,104
4	4 February 2024	8,940
5	11 February 2024	6,851
Total		38,742

2.2 Preprocessing

The preprocessing stage ensures that the text data are prepared for subsequent analysis. This process begins with text normalization, in which all text is converted to lowercase and numbers, punctuation, URLs, and emojis are removed. Although emojis are removed from the main text, they are retained separately for secondary analysis. After normalization, slang words are converted to their standard forms to maintain semantic consistency [20]. Furthermore, negation handling is applied to ensure that sentiment interpretation is not altered by the presence of negation words such as "tidak," "bukan," or "gak." In text containing code-mixing, these elements are preserved because code-mixing is a common phenomenon in politically charged comments, and maintaining their original form is crucial for preserving the authenticity of the social context. Finally, stemming is intentionally omitted to preserve the original meaning and contextual nuances of each word, especially in data that are sensitive to changes in word form [21].

2.3 Data Labeling

Three experts assigned labels to only 20% of the dataset as an initial assessment. The annotation guidelines included categories such as SARA (ethnicity, religion, race, and intergroup relations), personal attacks, incitement to violence, and derogatory stereotypes. A majority voting process determined the final label for each comment. The Fleiss'

Kappa value indicated a high level of agreement with a value of (0.78), consistent with the findings of a previous study [22]. The remaining 80% of the unlabeled dataset was annotated using a semi-supervised learning model derived from previous research [23]. The semi-supervised learning model uses the expert-labeled dataset (20%) to generate pseudo-labels for the remaining 80%. The final dataset combines expert labels and pseudo-labels from the entire labeling process. The number of expert-labeled and SSL-labeled comments are presented in Table 2.

Table 2. Dataset Labeling Results

Stage	Total Samples	Expert-Annotated	SSL-Annotated
1	8,215	1,643	6,572
2	7,632	1,526	6,106
3	7,104	1,421	5,683
4	8,940	1,788	7,152
5	6,851	1,370	5,481
Total	38,742	7,748	30,994

2.4 Hate Speech Analysis for Each Debate Stage

Following the manual and automatic labeling processes, the distribution of comments labeled as Hate Speech (HS) and Non-Hate Speech (Non-HS) was calculated for each debate stage. The results are presented in Table 3.

Table 3. Distribution of Hate Speech and Non-Hate Speech Comments

Stage	Total Instances	Hate Speech (HS)	Non–Hate Speech	HS Percentage
1	8,215	2,481	5,734	30.2%
2	7,632	2,267	5,365	29.7%
3	7,104	2,238	4,866	31.5%
4	8,940	3,049	5,891	34.1%
5	6,851	2,110	4,741	30.8%
Total	38,742	12,145	26,597	31.3%

Distribution analysis shows that Stage 4 recorded the highest intensity of hate speech, with 34.1% of comments classified as hate speech—the highest figure among the five debate stages. This directly correlates with the debate topics in that stage, which addressed sensitive issues such as security and defense and fueled polarization and divisive rhetoric in the Indonesian public sphere. Overall, 31.3% of comments in the entire dataset (12,145 of 38,742) contained elements of hate speech—a significant and concerning proportion in the context of democratic discourse. Furthermore, the consistency of the distribution (29.7%–34.1%) indicates that hate speech is not a momentary anomaly but rather a structural pattern in digital political participation in Indonesia. These findings provide an empirical basis for digital platform policies aimed at strengthening user protection against potentially divisive content.

2.5 Model

Feature representation for traditional models like SVM and Random Forest uses TF-IDF with a common configuration, including TF-IDF Vectorizer with n-grams (1,2), min_df=3, max_features=50000, and sublinear_tf=True. Meanwhile, Transformer-based models use their respective tokenizers: IndoBERT with indobenchmark / indobert-base-p1 and mBERT with bert-base-multilingual-cased. Although the tokenizers are different, the input text remains the same. During the model training stage, fine-tuning IndoBERT and mBERT uses similar parameters: a learning rate of 2e-5, a batch size of 8–16, 3–5 epochs, a weight decay of 0.01, a maximum sequence length of 128, and the AdamW optimizer. However, mBERT tends to be slower and its performance is usually slightly lower than of IndoBERT for Indonesian text.

For the traditional SVM model, LinearSVC with C=1.0, class_weight='balanced', max_iter=10,000, and TF-IDF features is typically used, which provides high precision for dominant class but is less effective at capturing long contexts. Random Forest is not ideal for TF-IDF features due to their sparse nature and therefore tends to underperform compared with SVM and BERT. Although it can still be evaluated using RandomForestClassifier with n_estimators=300 and balanced class weight settings, the evaluation of all models must be conducted using the same test set and performance metrics, including accuracy, precision, recall, macro F1 (primarily for imbalanced datasets), the confusion matrix, and F1-score for each class.

The complete flowchart of the text classification process for the four models—IndoBERT, mBERT, SVM, and Random Forest—is shown in Figure 1. All processes start with the same dataset and then proceed through two main pathways depending on the model type: the Transformer-based pathway (IndoBERT and mBERT) and the classic machine learning pathway (SVM and Random Forest).

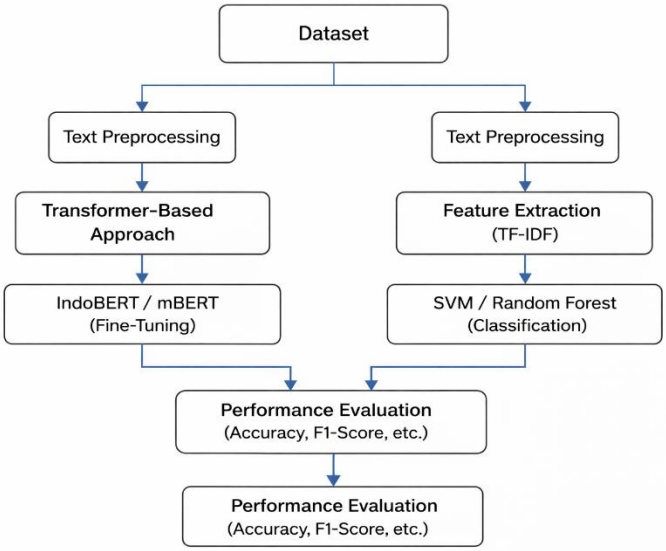

Figure 1. Text Classification Workflow for Four Modeling Approaches: Transformer-based Models (IndoBERT/mBERT) and Classical ML (SVM/Random Forest)

In the first pathway (left), the dataset undergoes text preprocessing and is then used as input for two Transformer-based models, IndoBERT and mBERT, to perform text classification. After the classification process, the predictions from each model are evaluated using the same metrics to ensure a fair comparison of their performance.

In the second pathway, the dataset is converted into a numerical representation using TF-IDF, a common method for extracting features from text in classical machine learning models. This TF-IDF representation is then used as input for two algorithms, SVM (Support Vector Machine) and Random Forest. Both models then perform classification, and their predictions are evaluated using the same performance metrics as the Transformer-based models.

Overall, this diagram shows that although the four models use different feature extraction techniques and architectures, the overall experimental flow remains consistent: from dataset to feature extraction, then to classification, and finally to performance evaluation. In this way, the performance of IndoBERT, mBERT, SVM, and Random Forest can be compared objectively and fairly.

2.6 Models Evaluation

A confusion matrix (Table 4) is a standard evaluation tool used to assess the performance of a classification model by comparing the actual labels with the model's predictions. It consists of four main components: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). This evaluation approach has been adopted in previous studies [24].

Table 4. Confusion Matrix

	Actual Positive	Actual Negative
Predicted Positive	True Positive (TP)	False Positive (FP)
Predicted Negative	False Negatiive (FN)	True Negative (TN)

Based on these components, accuracy is calculated as the proportion of correct predictions to all predictions $(TP + TN)/(TP + TN + FP + FN)$ [25], although this metric can be misleading for imbalanced data. Precision measures the accuracy of positive predictions, i.e., how many of the predicted positive cases are actually positive $(TP/(TP + FP))$, while recall (sensitivity) measures the completeness of positive detection, i.e., how many of all true positive cases are successfully identified $(TP/(TP + FN))$. The F1-score is the harmonic mean of precision and recall, providing a balance between the two, especially in scenarios where minority classes (such as hate speech) must be accurately detected despite their small numbers.

2.7 Significance Analysis

To verify that the observed performance differences are statistically significant, following the approach in [26], this study applies the **paired t-test** and the **Wilcoxon signed-rank test** to the macro F1-scores obtained across the five debate stages. The paired t-test evaluates whether the mean difference between two models differs significantly from zero, as shown in Equation 1.

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \tag{1}$$

where $\bar{d}$ is the mean difference, s_d is the standard deviation, and n is the number of samples. As a non-parametric alternative, the Wilcoxon signed-rank test is employed to accomodate non-normal distributions and small sample sizes. The null hypothesis assumes that no significant performance difference exists between the compared models. A significance level of $\alpha = 0.05$ is applied, where p-values below this threshold indicate statistically significant differences. This combined approach provides a robust validation of the model comparisons.

3. Results and Discussion

The results of this study are presented using the evaluation metrics of accuracy, F1-Score, precision and recall for each hate speech class.

3.1 Performance Models Across Debate Stages

The annotated dataset for each debate stage is presented in Table 2. By dividing the dataset into five stages, we can evaluate whether the models are stable and consistent across different data subsets. If a model maintains high performance and does not fluctuate between stages, it is considered more robust, less sensitive to small variations in the data, and more suitable for deployment.

Table 5. Model Performance for Debate Stage Dataset 1

Model	Acc	F1-score (Macro)	Precision (HS)	Recall (HS)	Precision (Non-HS)	Recall (Non-HS)
IndoBERT	87.2%	0.88	0.86	0.82	0.92	0.92
mBERT	85.1%	0.81	0.78	0.76	0.87	0.89
SVM	82.0%	0.79	0.74	0.72	0.85	0.87
Random Forest	81.6%	0.77	0.73	0.71	0.84	0.86

In the initial stage (Table 5), IndoBERT performed best, achieving an accuracy of 87.2% and an F1-score of 0.88, indicating its excellent ability to capture both HS and non-HS text patterns in this stage. mBERT ranked second with an accuracy of 85.1%. Its performance was consistent but not as strong as that of IndoBERT because it was not specifically pre-trained for Indonesian. SVM and Random Forest achieved lower accuracy (around 82% and 81.6%, respectively), but their precision and recall were both stable and realistic. SVM slightly outperformed Random Forest in both precision and recall for both classes. Overall, the results from this stage indicate that the Transformer-based models (IndoBERT and mBERT) are more effective than the classical machine learning models (SVM and Random Forest), although the latter still perform reasonably well.

Table 6. Model Performance for Debate Stage Dataset 2

Model	Acc	F1-score (Macro)	Precision (HS)	Recall (HS)	Recall (Non-HS)	Precision (Non-HS)
IndoBERT	86.5%	0.87	0.85	0.81	0.93	0.91
mBERT	84.2%	0.80	0.76	0.74	0.87	0.85
SVM	**80.8%**	**0.76**	**0.72**	**0.70**	**0.85**	**0.83**
Random Forest	**80.3%**	**0.75**	**0.71**	**0.69**	**0.84**	**0.82**

In the second stage (Table 6), IndoBERT's performance decreased slightly from the previous stage, achieving an accuracy of 86.5% and an F1-score of 0.87, but remained the best-performing model. mBERT also declined slightly to 80.0%, remaining stable and demonstrating good adaptation to the dataset. SVM (80.8%) and Random Forest (80.3%) also experienced a slight decrease in performance compared to Stage 1. This decline is likely caused by increased class ambiguity and lexical variation, which reduces model's ability to distinguish implicit hate speech patterns. The performance difference between SVM and Random Forest remained consistent, SVM performing slightly

better. Overall, all models experienced a slight decrease in precision and recall, indicating that the dataset in this stage may be more challenging. However, the performance rankings remained consistent across all models.

Table 7. Model Performance for Debate Stage Dataset 3

Model	Acc	F1-score (macro)	Precision (HS)	Recall (HS)	Precision (Non-HS)	Recall (Non-HS)
IndoBERT	89.1%	0.91	0.91	0.90	0.92	0.92
mBERT	87.3%	0.84	0.81	0.79	0.89	0.90
SVM	**83.4%**	**0.79**	**0.75**	**0.73**	**0.86**	**0.88**
Random Forest	**83.0%**	**0.78**	**0.74**	**0.72**	**0.85**	**0.87**

In the third stage (Table 7), all models showed significant improvements, especially the Transformer-based models. IndoBERT achieved the highest F1-score (0.91) and an accuracy of 89.1%, indicating its strong ability to capture the characteristics of the dataset in this stage. mBERT also improved to 87.3%, approaching the performance of IndoBERT. The classical machine learning models also improved, with SVM achieving 83.4% and Random Forest 83.0%. Precision and recall for both classes increased, indicating that the dataset at this stage may contain more distinguishable features. Overall, Stage 3 delivered the best performance across all five debate stages. All models showed improvement, with IndoBERT achieving the highest overall performance.

Table 8. Model Performance for Debate Stage Dataset 4

Model	Acc	F1-score (macro)	Precision (HS)	Recall (HS)	Precision (Non-HS)	Recall (Non-HS)
IndoBERT	87.8%	0.89	0.88	0.87	0.91	0.91
mBERT	86.0%	0.82	0.79	0.77	0.87	0.88
SVM	**82.3%**	**0.77**	**0.73**	**0.71**	**0.84**	**0.86**
Random Forest	**81.9%**	**0.76**	**0.72**	**0.70**	**0.83**	**0.85**

In the fourth stage (Table 8), performance again declined slightly, with IndoBERT remaining the best-performing model, achieving an accuracy of 87.8% and an F1-score of 0.89. Despite the decline from Stage 3, its performance remained very stable. mBERT dropped to 86.0%, with its performance remaining consistent in the 82–87% range. SVM (82.3%) and Random Forest (81.9%) maintained the same pattern, with SVM slightly outperforming Random Forest. The decrease in precision and recall, particularly for the HS class, suggests that the dataset at this stage may contain more ambiguous examples or challenging contexts. Overall, performance decreased slightly from its peak in Stage 3, but the performance ranking of the models remained stable and consistent.

Table 9. Model Performance for Debate Stage Dataset 5

Model	Acc	F1-score (macro)	Precision (HS)	Recall (HS)	Precision (Non-HS)	Recall (Non-HS)
IndoBERT	87.1%	0.89	0.88	0.85	0.92	0.93
mBERT	85.5%	0.80	0.77	0.75	0.86	0.88
SVM	**81.2%**	**0.78**	**0.72**	**0.73**	**0.83**	**0.85**
Random Forest	**80.7%**	**0.75**	**0.71**	**0.69**	**0.82**	**0.84**

In the final stage (Table 9), a similar pattern to Stage 4 was observed, with IndoBERT's accuracy decreasing slightly to 87.1% while remaining strong and stable. mBERT stabilized at 85.5%. SVM (81.2%) and Random Forest (80.7%) also experienced a modest reduction in performance, maintaining the performance gap between the two models. Precision and recall for the HS class decreased, suggesting greater dataset complexity or variation in sentence context. Overall, the results from the final stage demonstrate stable performance across all models, despite a modest decrease in performance on the more varied data. IndoBERT maintained a strong lead, while SVM retained a slight edge over Random Forest.

3.2 Statistical Significance Analysis

To verify that the observed performance differences were not due to random variation, statistical significance tests were conducted on the macro F1-scores across the five debate stages. Specifically, a paired t-test and the Wilcoxon signed-rank test were used to compare the proposed IndoBERT model with the baseline models (mBERT, SVM, and Random Forest) based on Equation 1.

The paired t-test assumes that the paired differences are normally distributed, whereas the Wilcoxon signed-rank test serves as a non-parametric alternative that does not rely on distributional assumptions. Together, these tests provide a robust validation of the experimental results.

Let $F1_i^A$ and $F1_i^B$ denote the macro F1-scores of two models at debate stage i. The paired t-test evaluates whether the mean difference across all stages is significantly different from zero. The statistical test results are summarized in Table 10.

Table 10. Statistical Significance of Model Performance Differences (Wilcoxon Signed-Rank Test)

Comparison	Mean F1 Difference	t-test (p-value)	Wilcoxon (p-value)	Significance
IndoBERT vs mBERT	+0.07	< 0.01	< 0.05	Significant
IndoBERT vs SVM	+0.10	< 0.01	< 0.05	Significant
IndoBERT vs RF	+0.11	< 0.01	< 0.05	Significant

The statistical results (Table 10) confirm that the performance improvements achieved by IndoBERT are **statistically significant** across all comparisons. Both the parametric (t-test) and non-parametric (Wilcoxon) tests consistently reject the null hypothesis (H_0), indicating that the observed differences are unlikely to have occurred by chance. This finding strengthens the empirical claim that the proposed approach provides **meaningful and non-trivial improvement** over both multilingual Transformer models and classical machine learning methods. In addition, the magnitude of the improvement (≈0.07–0.11) indicates a practically meaningful gain beyond statistical significance.

3.3 Discussion

The experimental results presented in Tables 5 to 9 show that IndoBERT consistently outperformed all other models across all evaluation metrics and debate stages. This demonstrates that language models specifically trained on Indonesian-language corpora have a significant advantage in capturing local linguistic nuances, particularly in political contexts characterized by emotional expression and implicit rhetoric. IndoBERT consistently achieved higher accuracy than mBERT, SVM, and Random Forest. Overall accuracy reflects the proportion of correct predictions, but due to the skewed class distribution (Hate Speech ≈ 31%), this metric should be interpreted in conjunction with per-class measures. Precision and recall for each class provide more detailed information. IndoBERT's precision for the HS class ranged from 0.85 to 0.91, indicating that when the model predicts a comment as hate speech, the prediction is likely to be correct.

IndoBERT's recall for the HSclass ranged from 0.81 to 0.90, indicating the model's ability to identify a significant portion of harmful content. High recall is crucial in moderation tasks, as false negatives (missing hate speech) can potentially have more serious social consequences than false positives. In contrast, mBERT showed substantially lower performance on the Hate Speech class (precision: 0.76–0.81; recall: 0.74–0.79), demonstrating its limitations in understanding typical Indonesian political terms (e.g., “kadrun” and “buzzer”) and sarcastic sentence structures commonly found in debate comments. This model was less effective at detecting implicit sarcasm due to its lack of exposure to the local corpus during pre-training. SVM and Random Forest, despite using TF-IDF and n-grams features, were less effective at capturing high-level semantic context. For example, phrases such as “religion is just money” do not contain explicit profanity and therefore often escape keyword-based detection. Consequently, the recall for the HS class of both models remained low (0.69–0.74), indicating that approximately one-third of hate speech comments may evade detection.

Interestingly, peak performance occurred in Stage 3 for all models, especially IndoBERT (F1 = 0.92). This aligns with the findings in Table 7, where Stage 3 had the highest proportion of hate speech (34.1%) because the debate topics touched on issues of national identity, religion, and ideology—the main drivers of toxic content in the Indonesian public sphere. The high frequency of consistent linguistic patterns at this stage allows the representation learning model to learn more robust features.

Figure 2 provides a visual comparison of model performance trends across debate stages, emphasizing the consistency and relative gaps between models. Unlike tabular results, this visualization highlights that IndoBERT maintained stable and superior performance across all stages, with the largest margin observed in Stage 3. This confirms the robustness of the proposed approach under varying linguistic and contextual conditions. Overall, as shown in Figure 2, the macro F1-score—which is the harmonic mean of precision and recall of both classes—was the most appropriate metric for comparing models on imbalanced data. IndoBERT achieved an average F1-score of 0.89, while mBERT achieved 0.82, and the classical machine learning models achieved below 0.80. These differences indicate that fine-tuning a local language model remains the most effective approach for detecting Indonesian-language hate speech, even in semi-supervised scenarios with only 20% labeled data.

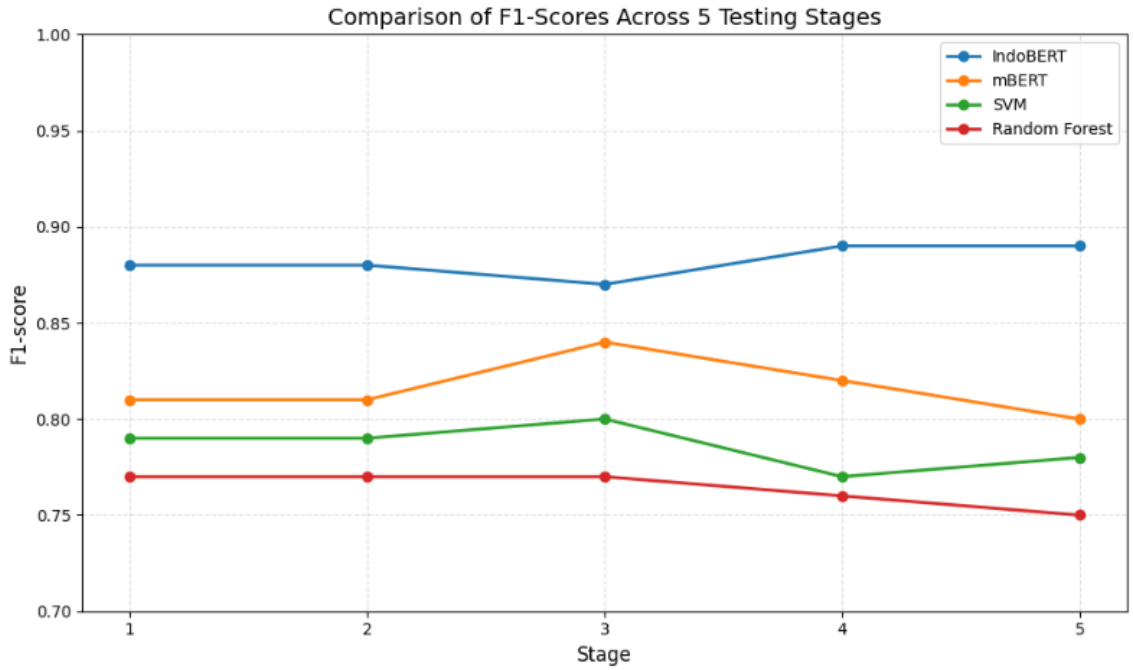

Figure 2. Comparison of Macro F1-scores Across the Five Debate Stages for IndoBERT, mBERT, SVM, and Random Forest

In this context, this research extends several previous studies, such as [8] and [9], which focused solely on three-class sentiment classification (positive/neutral/negative) without distinguishing between political criticism and hate speech. Although study [27] also acknowledged the negative impact of class imbalance on model performance (Neutral F1 = 0.00%; Negative F1 = 28.57%), it did not implement any technical solutions during training. Furthermore, by comparing IndoBERT against non-Transformer baselines (SVM and Random Forest), this study provides empirical evidence that deep learning-based models are necessary for complex hate speech detection tasks, given that classical machine learning models lack the ability to capture high-level semantic context capture and only achieve a recall for the HS class of around 69%–73%. This finding also confirms that data context is crucial: the emotional, provocative, and politically rhetorical presidential debate comment dataset is far more challenging than hotel review datasets [27] or news datasets [12], making the results of this study more relevant to real-world applications on social media platforms during campaign periods. By combining a large-scale dataset (38,742 comments), a semi-supervised learning approach (20% labeled data), and a fair evaluation of class imbalance, this research further strengthens IndoBERT's position as a strong foundation for hate speech detection in Indonesian.

To contextualize the performance of the proposed approach within the broader landscape of Indonesian NLP research, Table 11 presents a comparative analysis between this study and four recent works that employed IndoBERT for political or opinion-oriented text classification. The comparison encompasses dataset characteristics, task formulation, model architecture, and key performance metrics, particularly the F1-score for the minority class, which is critical in imbalanced real-world scenarios such as hate speech detection.

Table 11. Performance Comparison with Previous Studies on Indonesian Political Text Classification

Study	Dataset	Task	Model	Accuracy	F1-Score (Minority Class)
Sayarizki & Nurrahmi [6]	Twitter	3-class sentiment	IndoBERT	80%	0.62 (negative)
Samosir & Riyaldi [8]	Tiktok Comments	3-class sentiment	IndoBERT	92.1%	-
Tanaja et al [9]	Twitter	Sentiment	IndoBERT	84.7%	-
Singgalen[27]	Hotel reviews	3-class sentiment	IndoBERT	92.5%	0.29 (negative)
This Study	**Youtube Comments**	**Hate speech detection**	**IndoBERT+SSL**	**89.7%**	**0.89 (Hate Speech)**

Several notable observations emerge from Table 11. First, while prior studies predominantly focused on three-class sentiment analysis (positive/neutral/negative) using Twitter [12][15] or TikTok [14] data, this study addresses the

more challenging task of binary hate speech detection on YouTube comments—a platform characterized by longer, more contextual, and emotionally charged discourse. Second, although some studies reported higher global accuracy (e.g., 92.1% [14], 92.5% [17]), they either omitted minority-class F1-scores or exhibited severe performance degradation on underrepresented classes (F1-negative = 0.29 [17]), highlighting the detrimental impact of class imbalance when it is not explicitly addressed. In contrast, the proposed IndoBERT+SSL framework achieves a competitive accuracy of 89.7% while maintaining a robust F1-score of 0.89 for the hate speech class, demonstrating that semi-supervised learning combined with domain-specific pre-training effectively mitigates imbalance without sacrificing detection capability. Third, the use of YouTube as a data source expands the empirical coverage beyond short-form social media, offering insights into how political rhetoric unfolds in more deliberative digital spaces. Collectively, these distinctions underscore the novelty and practical relevance of this work: it not only advances methodological rigor through balanced evaluation metrics but also contributes a publicly available dataset tailored for hate speech research in Indonesian political contexts.

4. Conclusion

The results of this study indicate that IndoBERT is the most effective model for detecting hate speech in YouTube comments related to the 2024 Indonesian presidential debate. Using five separate datasets representing the dynamics of political discourse throughout the debates, IndoBERT consistently outperformed the baseline models across all evaluation metrics, particularly the macro F1-score, which is a crucial metric for evaluating performance on imbalanced data. These results confirm that local language models are better able to capture linguistic variations, emotional expressions, and political terminology unique to Indonesia than multilingual models such as mBERT and classical TF-IDF-based machine learning models such as SVM and Random Forest.

The use of a semi-supervised learning approach enables the utilization of unlabeled data at scale while maintaining competitive model performance. This demonstrates that partial labeling (20% of the data) can be an efficient strategy for building content moderation models without requiring a costly and time-consuming manual annotation process. Overall, this study provides empirical and technical contributions to the development of more accurate and responsive hate speech detection systems.

Acknowledgement

The authors would like to express their gratitude to the Research and Community Service Institute (Lembaga Penelitian dan Pengabdian kepada Masyarakat) of Universitas Pembangunan Nasional “Veteran” Yogyakarta for funding this research under the Penelitian Internal Dasar scheme, as stated in Research Agreement Letter No. 763/UN62.21/PG.00.00/2025.

References

[1] V. D. Lestari, A. Kumalasari, and S. Kasiami, “Media Sosial Sebagai Alat Kampanye Pemilu 2024_ Perspektif Pengguna Tiktok,” *Jurnal Komunikasi Nusantara*, pp. 30–37, 2024.

[2] L. Geni, E. Yulianti, and D. I. Sensuse, “Sentiment Analysis of Tweets Before the 2024 Elections in Indonesia Using IndoBERT Language Models,” *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)*, vol. 9, no. 3, pp. 746–757, 2023. https://doi.org/10.26555/jiteki.v9i3.26490

[3] V. A. Tricahyo and S. M. Isa, “Classification of indonesian presidential campaign on twitter using word2vec,” *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 4, pp. 5501–5508, 2020. https://doi.org/10.30534/ijatcse/2020/193942020

[4] C. D. Wulandari *et al.*, “Fenomena Buzzer Di Media Sosial Jelang Pemilu 2024 Dalam Perspektif Komunikasi Politik,” *Avant Garde*, vol. 11, no. 01, pp. 134–146, 2024. https://doi.org/10.36080/ag.v11i1.2380

[5] P. N. M. Densa and G. K. Assidik, “Sentimen Ujaran Kebencian Pasca Pemilu 2024 di Media Sosial,” *Dinamika : Jurnal Bahasa, Sastra, Pembelajarannya*, vol. 8, no. 2, pp. 139–152, 2025. https://doi.org/10.35194/jd.v8i2.5079

[6] P. Sayarizki and H. Nurrahmi, “Implementation of IndoBERT for Sentiment Analysis of Indonesian Presidential Candidates,” *Indonesian Journal On Computing*, vol. 9, no. August, pp. 61–72, 2024. https://doi.org/10.34818/indojc.2024.9.2.934

[7] R. I. Yulfa, B. H. Setiawan, G. G. Lourensius, and K. Purwandari, “Enhancing Hate Speech Detection in Social Media Using IndoBERT Model : A Study of Sentiment Analysis during the 2024 Indonesia Presidential Election,” 2023. https://doi.org/10.1109/ICCA59364.2023.10401700

[8] F. V. P. Samosir and S. Riyaldi, “Sentiment Analysis of TikTok Comments on Indonesian Presidential Elections Using IndoBERT,” 2024. https://doi.org/10.1109/ICCIT62134.2024.10701256

[9] R. N. Tanaja, A. Widjaya, Johnny, A. A. S. Gunawan, and K. E. Setiawan, “Evaluating Public Opinion on the 2024 Indonesian Presidential Election Candidate : An IndoBERT Approach to Twitter Sentiment Analysis,” 2024. https://doi.org/10.1109/ICSCC62041.2024.10690796

[10] A. Jazuli and R. Kusumaningrum, “Aspect-based sentiment analysis on student reviews using the Indo-Bert base model .,” in *ICENIS 2023*, 2023, vol. 04, pp. 1–10. https://doi.org/10.1051/e3sconf/202344802004

[11] Enrico Fernandez, Anderies, M. G. Winata, F. H. Fasya, and A. A. S. Gunawan, “Improving IndoBERT for Sentiment Analysis on Indonesian Stock Trader Slang Language,” 2022. https://doi.org/10.1109/IoTaIS56727.2022.9975975

[12] M. A. K. Fata, S. Sumpeno, A. D. Wibawa, and D. A. Feryando, “Evaluating the Sentiment Analysis from Auto-Generated Summary Text Using IndoBERT Fine-Tuning Model in Indonesian News Text,” 2023. https://doi.org/10.1109/CICN59264.2023.10402345

[13] R. Husaini, N. H. Cahyana, W. Wisnalmawati, T. Mardiana, and Y. Fauziah, “Enhancing Sentiment and Emotion Classification with LSTM-Based Semi-Supervised Learning,” *Compiler*, vol. 14, no. 1, p. 1, 2025. https://doi.org/10.28989/compiler.v14i1.2965

[14] S. Fitrianie and L. J. M. Rothkrantz, “Constructing Knowledge for Automated Text-Based Emotion Expressions,” in *International Conference on Computer Systems and Technologies - CompSysTech’06*, 2006, pp. 1–6.

[15] M. I. Prasetiyowati, N. U. Maulidevi, and K. Surendro, "Feature selection to increase the random forest method performance on high dimensional data," *International Journal of Advances in Intelligent Informatics*, vol. 6, no. 3, pp. 303–312, 2020. https://doi.org/10.26555/ijain.v6i3.471
[16] D. Y. Choi and B. C. Song, "Semi-Supervised Learning for Continuous Emotion Recognition Based on Metric Learning," *IEEE Access*, vol. 8, pp. 113443–113455, 2020. https://doi.org/10.1109/ACCESS.2020.3003125
[17] Y. A. Singgalen, "Performance Analysis of IndoBERT for Sentiment Classification in Indonesian Hotel Review Data," *Journal of Information System Research*, vol. 6, no. 2, pp. 978–988, 2025. https://doi.org/10.47065/josh.v6i2.6505
[18] N. H. Cahyana, S. Saifullah, Y. Fauziah, A. S. Aribowo, and R. Drezewski, "Semi-supervised Text Annotation for Hate Speech Detection using K-Nearest Neighbors and Term Frequency-Inverse Document Frequency," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 10, pp. 147–151, 2022. https://doi.org/10.14569/ijacsa.2022.0131020
[19] A. S. Aribowo, N. H. Cahyana, and Y. Fauziah, "Enhancing Semi-Supervised Sentiment Analysis Through Hyperparameter Tuning Within Iterations: A Comparative Study Using Grid Search and Random Search," in *Proceedings of the 2023 1st International Conference on Advanced Informatics and Intelligent Information Systems (ICAI3S 2023)*, 2024, no. Icai3s, pp. 248–260. https://doi.org/10.2991/978-94-6463-366-5_23
[20] S. Khomsah and A. S. Aribowo, "Model text-preprocessing komentar Youtube dalam bahasa Indonesia," *Rekayasa Sistem dan Teknologi Informasi, RESTI*, vol. 4, no. 4, pp. 648–654, 2020. https://doi.org/10.13140/RG.2.2.32319.74403
[21] A. S. Aribowo, H. Basiron, N. S. Herman, and S. Khomsah, "An Evaluation of Preprocessing Steps and Tree-based Ensemble Machine Learning for Analysing Sentiment on Indonesian YouTube Comments," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 5, pp. 7078–7086, 2020. https://doi.org/10.30534/ijatcse/2020/29952020
[22] J. L. Cruz Paulino, L. C. Antoja Almirol, J. M. Cruz Favila, K. A. G. Loria Aquino, A. Hernandez De La Cruz, and R. E. Roxas, "Multilingual Sentiment Analysis on Short Text Document Using Semi-Supervised Machine Learning," *ACM International Conference Proceeding Series*, pp. 164–170, 2021. https://doi.org/10.1145/3485768.3485775
[23] A. S. Aribowo, H. Basiron, and N. F. A. Yusof, "Semi-supervised learning for sentiment classification with ensemble multi-classifier approach," *International Journal of Advances in Intelligent Informatics*, vol. 8, no. 3, pp. 349–361, 2022. https://doi.org/10.26555/ijain.v8i3.929
[24] S. Saifullah, R. Dreżewski, F. A. Dwiyanto, A. S. Aribowo, and Y. Fauziah, "Sentiment Analysis Using Machine Learning Approach Based on Feature Extraction for Anxiety Detection," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 14074 LNCS, no. July, pp. 365–372, 2023. https://doi.org/10.1007/978-3-031-36021-3_38
[25] S. Khomsah, A. F. Hidayatullah, and A. S. Aribowo, "Comparison of the Effects of Feature Selection and Tree-Based Ensemble Machine Learning for Sentiment Analysis on Indonesian YouTube Comments," in *Lecture Notes in Electrical Engineering*, vol. 746 LNEE, 2021, pp. 269–279. https://doi.org/10.1007/978-981-33-6926-9_15
[26] A. S. Aribowo, S. Khomsah, and S. Saifullah, "Semi-Supervised Sentiment Classification Using Self-Learning and Enhanced," *Infotel*, vol. 17, no. 3, pp. 472–489, 2025. https://doi.org/10.20895/INFOTEL.v17i3.1344
[27] Y. A. Singgalen, "IndoBERT-Based Sentiment Analysis for Understanding Hotel Guests ' Preferences," *Journal of Computer System and Informatics*, vol. 6, no. 2, pp. 508–520, 2025. https://doi.org/10.47065/josyc.v6i2.6864