Original Article

Available online at
www.heca-analitika.com/ijds

Infolitika Journal of Data Science

Vol. 2, No. 2, 2024

Performance Assessment of Machine Learning and Transformer Models for Indonesian Multi-Label Hate Speech Detection

Ricky Bagestra [1], Alim Misbullah [1,*], Zulfan Zulfan [1], Rasudin Rasudin [1], Laina Farsiah [1] and Sri Azizah Nazhifah [1]

[1] Department of Informatics, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, Banda Aceh 23111, Indonesia; rickybages@gmail.com (R.B.); misbullah@usk.ac.id (A.M.); zulfan.abdullah@usk.ac.id (Z.Z.); rasudin@usk.ac.id (R.R.); lainafarsiah@usk.ac.id (L.F.); sriazizah07@usk.ac.id (S.A.N.)

* Correspondence: misbullah@usk.ac.id

Article History

Received 11 September 2024
Revised 14 November 2024
Accepted 20 November 2024
Available Online 28 November 2024

Keywords:
Social media platform
Hate speech
Support Vector Machine
Naive Bayes
IndoBERT

Abstract

Hate speech, characterized by language that incites discrimination, hostility, or violence against individuals or groups based on attributes such as race, religion, or gender, has become a critical issue on social media platforms. In Indonesia, unique linguistic complexities, such as slang, informal expressions, and code-switching, complicate its detection. This study evaluates the performance of Support Vector Machine (SVM), Naive Bayes, and IndoBERT models for multi-label hate speech detection on a dataset of 13,169 annotated Indonesian tweets. The results show that IndoBERT outperforms SVM and Naive Bayes across all metrics, achieving an accuracy of 93%, F1-score of 91%, precision of 91%, and recall of 91%. IndoBERT's contextual embeddings effectively capture nuanced relationships and complex linguistic patterns, offering superior performance in comparison to traditional methods. The study addresses dataset imbalance using BERT-based data augmentation, leading to significant metric improvements, particularly for SVM and Naive Bayes. Preprocessing steps proved essential in standardizing the dataset for effective model training. This research underscores IndoBERT's potential for advancing hate speech detection in non-English, low-resource languages. The findings contribute to the development of scalable, language-specific solutions for managing harmful online content, promoting safer and more inclusive digital environments.

1. Introduction

In recent years, social media has become deeply embedded in the daily lives of Indonesian society. Technological advancements have transformed communication into a vital skill in the digital era, reshaping how individuals interact, exchange ideas, and consume information [1]. Platforms like Twitter and Facebook, which facilitate real-time interactions and content sharing, have become key arenas for public discourse. However, the unregulated nature of user-generated content on these platforms has led to significant challenges, particularly the proliferation of hate speech [2, 3]. Defined as any form of communication that denigrates an individual or group based on characteristics such as race, religion, gender, or ethnicity, hate speech poses profound social and legal concerns [4]. In the Indonesian context, hate speech is not only a moral issue but also a legal offense with serious repercussions for individuals and communities.

In Indonesia, the issue of hate speech on digital platforms has become particularly significant due to the country's rapid digitalization and the widespread use of social

media. With over 200 million internet users, Indonesia ranks among the largest social media markets in the world, making it a fertile ground for both positive engagement and harmful content, including hate speech [5]. The case of I Gede Ari Astina, better known as Jerinx, serves as a prominent example, underscoring the severity of the problem and the tangible consequences of harmful online interactions [6]. These instances illustrate how hate speech can escalate from digital platforms to real-world conflicts, exacerbating social divisions and fostering hostility.

Addressing hate speech in the digital domain requires sophisticated and scalable solutions. Automated detection systems leveraging Natural Language Processing (NLP) and machine learning have emerged as effective tools for identifying and mitigating harmful content. These systems are designed to classify textual data with high accuracy, enabling proactive measures to curtail the spread of hate speech, protect users, and support authorities in managing online spaces [7, 8]. While traditional machine learning methods like Support Vector Machines (SVM) and Naive Bayes offer a foundation for such systems, the advent of deep learning models, particularly pre-trained models like IndoBERT, has revolutionized the field by enhancing contextual understanding and detection accuracy in non-English languages [9].

Despite these advancements, several challenges remain. Dataset imbalance is a critical issue, as hate speech categories often exhibit skewed distributions, with minority classes being underrepresented. Techniques such as BERT-based data augmentation have been introduced to address this imbalance, expanding the dataset to ensure equitable representation of all categories and improving model performance in recognizing nuanced hate speech patterns [10]. Furthermore, preprocessing steps such as text normalization, stemming, and tokenization are pivotal in optimizing datasets for machine learning tasks, particularly in morphologically rich languages like Indonesian [11, 12].

This study aims to evaluate and compare three methods for detecting hate speech: SVM, Naive Bayes, and IndoBERT. These methods represent both traditional and modern approaches, helping to understand their strengths and weaknesses and how well they work in identifying hate speech in the Indonesian language. By using techniques like preprocessing, hyperparameter tuning, and data augmentation, this research seeks to improve the models' ability to detect hate speech accurately, even when dealing with unevenly distributed data. The focus on Indonesian highlights the importance of building tools that consider the specific language and cultural features of non-English-speaking communities.

The results of this research can benefit many groups. For researchers, comparing traditional and deep learning models adds to the knowledge about which methods work best for hate speech detection. For social media platforms and authorities, these findings can help create better systems to reduce harmful content and make online spaces safer. By providing practical solutions, this study supports efforts to address hate speech and promote a more respectful and inclusive digital environment.

2. Related Works

Research focusing on hate speech detection in the Indonesian language has received significant attention, addressing various methodological and contextual challenges. A prominent study by Ibrohim & Budi [13] employed the Random Forest Decision Tree (RFDT) method, achieving an accuracy of 77.36%. This research provided a foundational multi-label dataset, which serves as a crucial resource for subsequent studies. However, RFDT struggles with capturing the nuanced and contextual nature of hate speech, limiting its applicability in complex scenarios.

In another relevant work by Wenando & Fuad [14], four algorithms—Naive Bayes, SVM, Decision Tree, and Random Forest—were evaluated. Naive Bayes emerged as the most effective model, with an accuracy of 88.57%. Despite its simplicity and effectiveness, Naive Bayes is limited in understanding deeper contextual relationships, making it suboptimal for tasks involving subtle linguistic features.

In the domain of sentiment analysis, studies conducted by Devlin et al. [15] highlight the potential of deep learning models. Fine-tuned multilingual BERT and IndoBERT achieved competitive results, with IndoBERT reaching an accuracy of 82.2%. While this demonstrates the utility of IndoBERT in the Indonesian language context, the focus on sentiment analysis rather than hate speech leaves a gap for further exploration.

Building on this, Dharmawan et al. [16] showcased the efficacy of IndoBERT in classifying hate speech, attaining an accuracy of 89.52%. While promising, the study did not benchmark IndoBERT against traditional methods such as SVM or Naive Bayes, limiting insights into comparative performance.

In another study by Yazid et al. [17], Naive Bayes was combined with N-Gram feature extraction to detect hate speech in a multi-label dataset. The resulting accuracy of

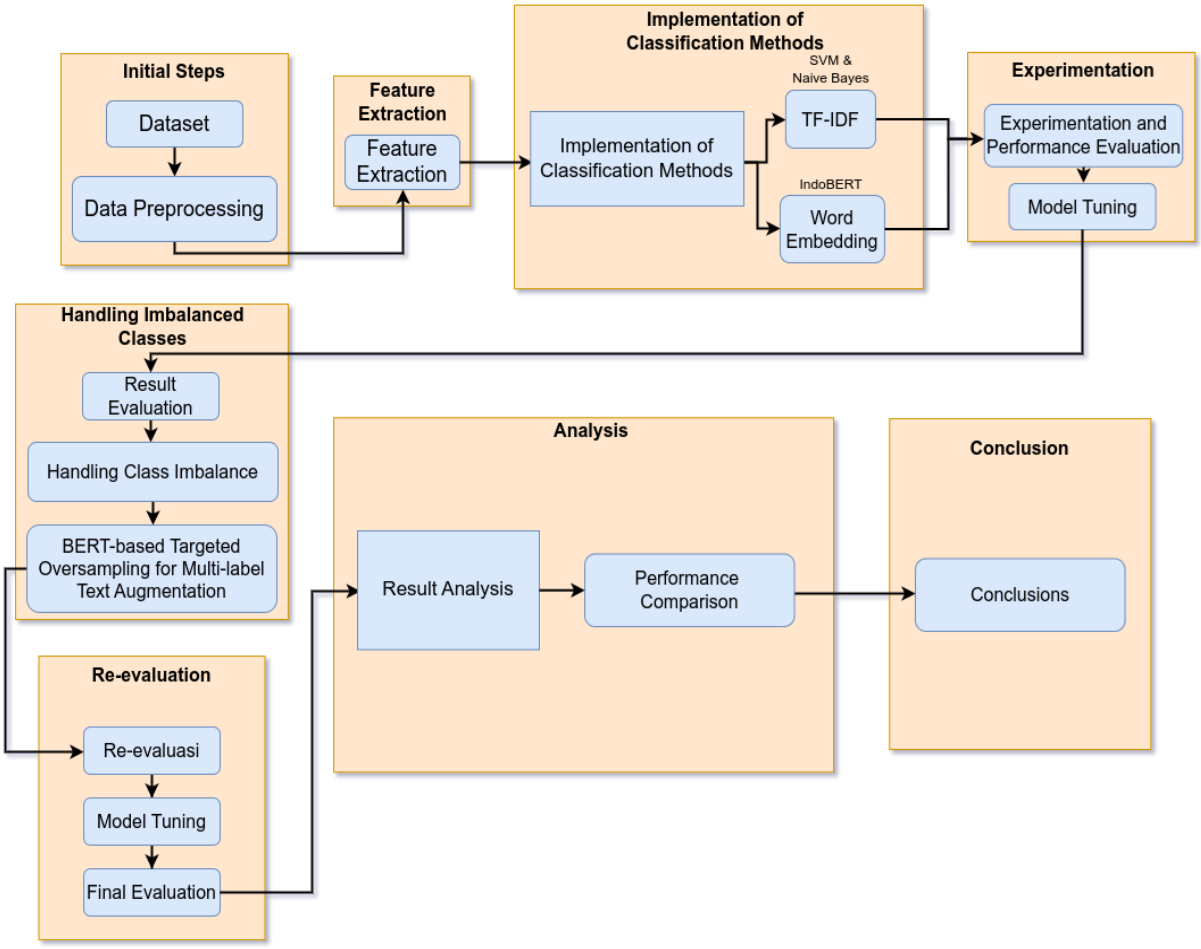

Figure 1. Overview of the methodology for hate speech detection.

64.95% revealed that method combinations might not always lead to performance improvements, underscoring the importance of tailoring methodologies to specific datasets.

A notable exploration of imbalance issues in hate speech detection was conducted Sanya and Suadaa [18], who explored hate speech detection in Indonesian online comments, highlighting the impact of imbalanced datasets on model performance. Fine-tuned IndoBERT outperformed SVM in extreme imbalance scenarios while combining SVM with SMOTE yielded the best results. However, BERT-based data augmentation, effective in similar tasks, remains unexplored for Indonesian datasets.

Pre-trained transformer models like IndoBERT have gained traction due to their contextual understanding and robustness in text classification tasks [19]. Vaswani et al. [20] introduced the transformer architecture, which underpins IndoBERT, enabling advanced capabilities in sequence modeling and context retention. However, comparative studies exploring IndoBERT alongside classical methods on multi-label datasets are scarce.

Additionally, research on optimizing model performance through hyperparameter tuning and ensemble methods has shown promise in hate speech detection. Pen et al. [21] demonstrated that a combination of traditional and deep learning models could yield improved results, though such approaches remain underexplored in the Indonesian context.

Despite these advancements, no single study has comprehensively compared SVM, Naive Bayes, and IndoBERT within the scope of multi-label hate speech detection in Indonesia. This research addresses this gap by evaluating these methods on a standard dataset, incorporating advanced preprocessing and augmentation techniques, and analyzing the impact of hyperparameter tuning. By bridging these gaps, the study aims to contribute a nuanced understanding of the optimal approaches for detecting hate speech in the Indonesian language.

3. Materials and Methods

The methodology for this study, as shown in Figure 1, systematically tackles text classification model

implementation, evaluation, and refinement while addressing challenges like imbalanced datasets. It begins with Initial Steps, including Data Preprocessing to clean the dataset and Feature Extraction using TF-IDF and IndoBERT embeddings for text transformation.

Next, the Implementation of the Classification Methods phase applies SVM and Naive Bayes algorithms, with Experimentation focusing on model tuning and performance evaluation. To handle imbalanced data, a Handling Imbalanced Classes stage uses BERT-based text augmentation to improve underrepresented class detection. The Analysis phase compares model performances, and if needed, the Re-evaluation stage fine-tunes models for better results. Finally, the Conclusion summarizes the findings and evaluates the methods used.

3.1. Dataset

The dataset used in this study is sourced from the research by Ibrohim & Budi [13], comprising comments collected from Twitter. It was specifically gathered to support the detection of hate speech and abusive language. Each comment is labeled to enable effective classification and analysis.

3.2. Data Preprocessing

Data preprocessing is an essential initial step in data analysis aimed at ensuring data cleanliness, accuracy, and readability prior to further analysis. This process involves techniques such as data cleaning, normalization, and dimensionality reduction, tailored to the specific requirements of the analysis. Key pre-processing techniques include noise removal, which eliminates irrelevant information such as special characters or symbols that may interfere with analysis, and case folding, which converts all text to lowercase to ensure consistency [22]. Additionally, slang word translation replaces informal words with their formal equivalents to enhance semantic clarity while stemming reduces words to their root forms using tools like Satrawi for the Indonesian language [23]. Finally, stopword removal eliminates common words that are less relevant, allowing a focus on critical keywords. Effective pre-processing is crucial for producing accurate and reliable analysis results

3.3. Feature Extraction

Feature extraction is a crucial step in text classification, transforming raw text data into numerical representations. This study employs two main methods: TF-IDF (Term Frequency-Inverse Document Frequency) and word embedding with IndoBERT, tailored to the characteristics of the classification models.

TF-IDF calculates word weights based on their frequency in a document and their rarity across the entire corpus. This method is used for SVM and Naive Bayes models as it generates vector representations suitable for these approaches.

Word embedding with IndoBERT, an alternative to TF-IDF, produces contextual vector representations using transformer architectures. IndoBERT, specifically trained for the Indonesian language, captures the relationships between words within a sentence, providing more informative features for deep learning models [24].

TF-IDF is not used with IndoBERT, as IndoBERT employs an internal context-based embedding mechanism, which is more appropriate than the static representations generated by TF-IDF. Therefore, TF-IDF is applied to SVM and Naive Bayes, while IndoBERT relies on its internal contextual embeddings [25].

3.4. Model Training

The model training phase is the core of this research, focusing on classifying input text into hate speech or non-hate speech categories. This process involves the application of three distinct methods: SVM, Naive Bayes, and IndoBERT. Each method has been tailored to maximize performance based on its unique characteristics and compatibility with feature extraction techniques.

3.4.1. Support Vector Machine (SVM)

SVM is one of the most widely developed classification methods today. The fundamental concept of this method is to maximize the margin of the hyperplane that separates a dataset [26]. SVM operates by finding an optimal hyperplane within a feature space to distinguish between two classes. The hyperplane selected maximizes the distance from the nearest data points on either side. The basic equation of SVM represents the hyperplane separating positive and negative classes, as shown in Equation 1:

$$f(x) = w^T x + b = 0 \quad (1)$$

where w represents the weight vector, x represents the input vector, and b is the bias term.

SVM aims to maximize the margin, which is the distance between the nearest data points of the two classes and the hyperplane. This goal is achieved by solving an optimization problem, as stated in Equation 2:

$$min\, w \cdot b\, \frac{1}{2}||w||^2 \quad (2)$$

where $\|w\|$ is the norm of the weight vector $w \frac{1}{2}\|w\|^2$ w and bias term.

Additionally, there are constraints to ensure that all data points in the two classes are correctly separated by the hyperplane. These constraints can be expressed in Equation 3:

$$y_i(w^T x_i + b) \geq 1, \forall_i \quad (3)$$

where y_i represents the class label of the i i-th data point, with $y_i = +1$ for one class and $y_i = -1$ for other classes, x_i is the feature vector of the i i-th data point, and $w^T x_i + b$ is the predicted value for the i-th data, which must match its label for the data to be correctly classified. This constraint ensures that all data points from the two classes are positioned on the correct side of the hyperplane, with a minimum distance of 1 from the hyperplane for the nearest data points.

3.4.2. Naive Bayes

After feature extraction using TF-IDF, the next step is to train the model using Naive Bayes. The Naive Bayes classification method has been widely applied in research domains that aim to classify large datasets. An international journal highlights that Bayes' Theorem is a concept of probability rules, determining true and false probabilities to derive additional information or knowledge [27]. Bayes' Theorem can be expressed in Equation 4:

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)} \quad (4)$$

Where $P(C|X)$ is the probability of class C given data X, $P(X|C)$is the probability of observing X given class C, $P(C)$ is the prior probability of class C, and $P(X)$ is the probability of the feature X.

3.4.3. IndoBERT

IndoBERT, a pre-trained language model developed specifically for the Indonesian language, is a variant of the BERT (Bidirectional Encoder Representations from Transformers) architecture that has been fine-tuned using Indonesian corpora. Leveraging its ability to understand linguistic nuances and contextual information, IndoBERT outperforms traditional machine learning methods in tasks such as text classification, sentiment analysis, and hate speech detection. Unlike generic models trained on multilingual datasets, IndoBERT is optimized for the syntax and semantics unique to Indonesian, enabling higher accuracy in processing complex text structures and colloquial expressions. Its effectiveness in hate speech detection stems from its ability to encode context-dependent features, making it highly suitable for addressing challenges in multi-label classification tasks.

In this research, IndoBERT's performance was optimized using specific configurations and techniques:

- Optimizer (Adam): The Adam optimizer, which combines RMSprop and Stochastic Gradient Descent (SGD), was utilized for its computational efficiency, ease of implementation, and ability to handle noisy gradients effectively. Adam is particularly advantageous in dealing with high-dimensional data, as it adapts learning rates for individual parameters during training.
- Loss (Categorical Crossentropy): This loss function was employed to address the binary classification nature of the problem, with target values in the set {0,1}. Categorical Crossentropy measures the dissimilarity between predicted and actual probability distributions, ensuring robust model convergence.
- Validation Split: To evaluate model performance, the dataset was divided into 90% for training and 10% for validation. This approach ensures that the model is assessed on unseen data, reducing the risk of overfitting. Accuracy was calculated by comparing the training error, which reflects classification performance on training data, and test error, determined using the separate validation set.
- Batch Size (32): A batch size of 32 was selected, allowing the model to process 32 samples per iteration. This size balances computational efficiency with stable gradient updates during training.
- Epochs: The number of training epochs was adjusted dynamically to achieve optimal performance. Training continued until the model's results aligned with the desired performance metrics, ensuring that the model effectively learned the patterns in the dataset.

3.5. Class Imbalance Handling

To address class imbalance in the dataset, a data augmentation technique based on BERT is used. Augmentation is performed by leveraging Masked Language Modeling (MLM) to generate variations of data from the minority class. This technique aims to balance the data distribution across classes, allowing the model to better detect both the majority and minority classes [28].

3.6. Evaluation

The evaluation of this model aims to assess the accuracy of its performance. In this case, the confusion matrix and

Table 1. Parameters before and after tuning.

Parameter	Before Tuning	After Tuning
SVM		
Kernel	RBF	RBF
C	1	10
Gamma	Scale	Scale
Class_weight	None	0
Probability	0	1
Random State	42	42
Naive Bayes		
Alpha	1	0.1
Random State	42	42
IndoBERT		
Model Architecture	indolem/indobert-base-uncased	indobenchmark/indobert-base-p1
Learning Rate	0	0
Batch Size	16	16
Epochs	5	5
Maximum Sequence Length	512	512
Random State	42	26092020

an accuracy table are used, along with the precision of each model being examined. After testing is performed on the training data, several classes from the test data are predicted, referred to as class predictions. These predicted classes, originally derived from the test data, are then concealed to allow for the display and calculation of the accuracy, precision, recall, and F1-score [29].

4. Results and Discussion

In the performance comparison of the three classification models used in this study, namely SVM, Naive Bayes, and IndoBERT, the performance of each model was evaluated using metrics such as accuracy, F1-score, recall, and precision, both before and after the parameter tuning process. Table 1 highlights the hyperparameter configurations for SVM, Naive Bayes, and IndoBERT models, comparing their settings before and after tuning. For SVM, the kernel remained RBF (Radial Basis Function), suitable for capturing non-linear patterns, while the C parameter was increased from 1 to 10, encouraging the model to fit the training data more strictly. Additionally, the class_weight was adjusted from "None" to 0, addressing imbalanced classes by giving more importance to underrepresented data. The probability parameter was also enabled after tuning, allowing the model to provide confidence scores for its predictions. Other parameters, such as gamma (set to "Scale") and random state (42), were retained, ensuring stable and reliable results.

For Naive Bayes, the tuning process focused on reducing the alpha parameter from 1 to 0.1, enabling finer adjustments in probability smoothing and improving the model's ability to capture nuanced patterns in text. Meanwhile, IndoBERT underwent a significant change in its architecture, transitioning from indole/indoor-base-uncased to the more optimized indobenchmark/indoor-base-p1, which provides better contextual embeddings for Indonesian text. Other parameters, such as batch size (16), maximum sequence length (512), and epochs (5), remained unchanged, ensuring efficient training while handling longer text sequences. However, the random state was altered from 42 to 26092020, introducing variability to potentially enhance training diversity.

These tuning adjustments demonstrate a targeted approach to optimizing model performance. SVM and Naive Bayes were refined to handle imbalanced data and improve generalization, while IndoBERT benefited from a more advanced pre-trained architecture, aligning with the task's requirements. Together, these modifications enhance the models' capabilities, ensuring more robust and accurate predictions.

4.1. Preprocessing Results

The data underwent a series of careful and systematic pre-processing steps. This pre-processing phase is a crucial stage in data analysis, aimed at improving data quality, ensuring consistency, and preparing the data for optimal use in further analysis. The pre-processing steps performed in this study including lower case handling, remove unnecessary char, remove non-alphanumeric, and slang word normalization as shown in Table 2.

The next step involved removing stopwords, which are common words like "dan" (and), "yang" (that), and "di" (in) that do not carry significant meaning for analysis. Subsequently, stemming was performed to reduce words to their root form by removing affixes, thereby standardizing variations of words with similar meanings and reducing the vocabulary dimension in the document. In addition, excessive spaces were removed to ensure the

Table 2. Pre-processing result.

Process	Before	After
Lower Case	Berhasil Penjarakan Ahok, Alumni 212 Pendukung Rizieq Akan Lengserkan Jokowi Usai Lebaran	berhasil penjarakan ahok, alumni 212 pendukung rizieq akan lengserkan jokowi usai lebaran
Remove Unnecessary Char	user wakakakaka aku sudah mengalaminya \xf0\x9f\x98\x82\xf0\x9f\x98\x82\xf0\x9f\x98\x82\xf0\x9f\x98\x82 tetiba muncul grub wa a.n bani abd hadi a.k.a mbahku dan seketika aku mek kenal beberapa orang saja \xf0\x9f\xa4\xa3\xf0\x9f\xa4\xa3\xf0\x9f\xa4\	wakakakaka aku sudah mengalaminya tetiba muncul grub wa an bani abd hadi aka mbahku dan seketika aku mek kenal beberapa orang saja
Remove Non-Alphanumeric	Kelakuan hom* jaman now, ngent*t aja sambil live di Blued #g*yvid #gvid #lokalhangat URL"	Kelakuan hom* jaman now ngent*t aja sambil live di blued g*yvid gvid lokalhangat url
Slang Word Normalization	Klo gk mau macet pulng aj lo ke kampung sono.. Dsar orng udik.. Masih nyri makan sma buang tai aj lelaguan... #anakjakartaasikasik'	Kalau gak mau macet pulang aja lo ke kampung sana dasar orang udik masih nyari makan sama buang tai aja lelaguan anak jakarta asik asik

Table 3. Performance evaluation comparison before and after tuning.

No.	Methods	Metrics	Before Tuning (%)	After Tuning (%)	Difference (%)
1	Support Vector Machine (SVM)	Accuracy	61	62	+1
		F1-Score	66	69	+3
		Recall	57	63	+6
		Precision	82	78	-4
2.	Naive Bayes	Accuracy	54	56	+2
		F1-Score	59	66	+7
		Recall	51	60	+9
		Precision	79	75	-4
3.	IndoBERT	Accuracy	61	91	+30
		F1-Score	72	84	+12
		Recall	73	83	+10
		Precision	71	86	+15

text format was more structured, and missing data were addressed either by deletion or handling empty values to maintain the dataset's quality.

4.2. Performance Evaluation Comparison Before and After Tuning

The evaluation results before the tuning process showed significant performance differences among the three methods tested for hate speech detection. Both SVM and IndoBERT demonstrated highly competitive performance, with identical accuracy and F1-Score values of 61% and 72%, respectively, as shown in Table 3. However, there were slight differences in recall and precision, offering valuable insights into the characteristics of each model.

Before tuning, there were significant differences in the hate speech detection performance among the three methods. SVM and IndoBERT outperformed Naive Bayes, with accuracy and F1-Score reaching 61% and 72%, respectively. SVM had a higher precision (82%) but lower recall (57%) compared to IndoBERT (precision 71%, recall 73%), which was more effective in detecting hate speech. Naive Bayes recorded high precision (79%) but low recall (51%), indicating a tendency to miss many hate speech cases. After tuning, performance improved significantly, especially for IndoBERT, with accuracy rising to 91%, F1-Score increasing to 84%, and precision reaching 86%. This improvement was supported by the use of a pre-trained architecture (indobenchmark/indobert-base-p1), the Adam optimizer with a low learning rate (1e-5), and a batch size of 16. SVM showed improvements in accuracy and F1-Score with the RBF kernel and parameter C (10.0), though precision slightly decreased. Naive Bayes with alpha 0.1 also saw an increase in F1-Score from 59% to 66%. Overall, tuning successfully enhanced the performance of all three models. IndoBERT showed the greatest improvement, while SVM and Naive Bayes remain relevant, especially for scenarios with limited computational resources or the need for model interpretability.

4.3. Impact of Handling Class Imbalance

To address the class imbalance in the multi-label hate speech dataset, this study employed a BERT-based augmentation technique using masked language modeling (MLM) to generate synthetic data samples. This technique enriches data variation without significantly altering the original context, such as changing *"Orang ini sangat bodoh"* to *"Orang tersebut sangat bodoh."* After augmentation, the dataset size increased from 13,114 to 20,759 samples (58.29%), improving the class distribution, particularly for the HS (Hate Speech) label, with the proportion rising from 42.34% to 64.77% (an

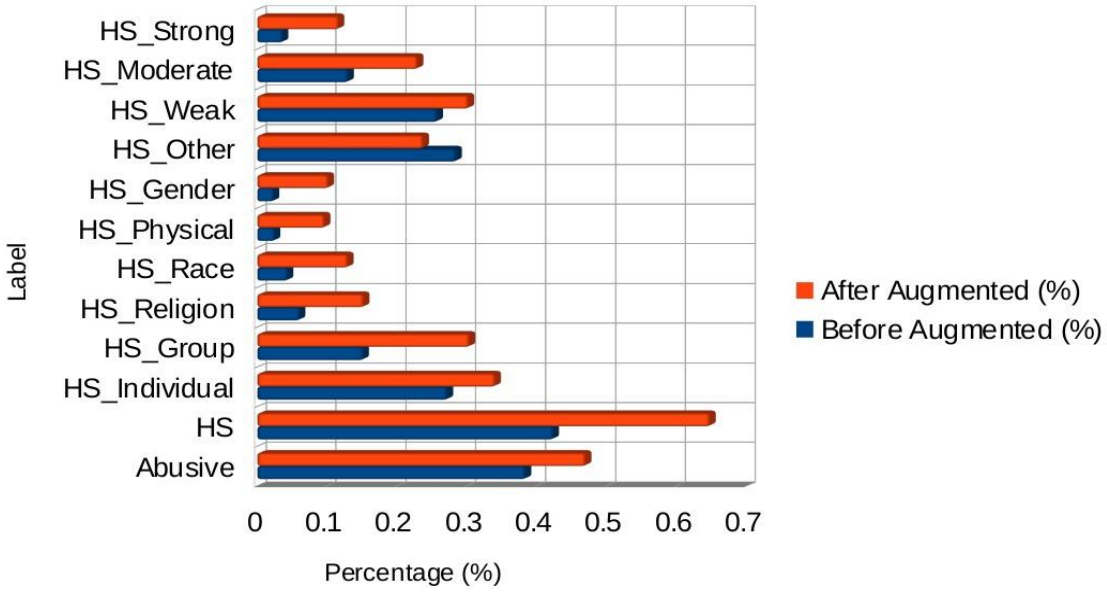

Figure 2. Label distribution changes before and after BERT-based data augmentation.

Table 4. Comparison of method performance before and after class imbalance handling.

No.	Methods	Metrics	Before Handling (%)	After Handling (%)	Difference (%)
1.	SVM	Accuracy	62	76	+14
		F1-Score	69	90	+21
		Recall	63	89	+26
		Precision	78	92	+14
2.	Naive Bayes	Accuracy	56	51	-5
		F1-Score	66	80	+14
		Recall	60	76	+16
		Precision	75	86	+11
3.	IndoBERT	Accuracy	91	93	+2
		F1-Score	84	91	+7
		Recall	83	91	+8
		Precision	86	91	+5

absolute increase of 22.43%). This technique effectively added diversity to the minority class, allowing the model to learn more effectively without drastically altering the data distribution.

In Figure 2, the percentage of the HS_Other label in the class distribution after augmentation decreased due to the increase in sample numbers for other labels, such as Abusive, HS_Individual, and HS_Religion. Although the number of HS_Other samples remained the same or increased, its proportion decreased. This reduction indicates that the data augmentation successfully balanced the class distribution, enhanced the representation of minority labels, and is expected to improve the model's accuracy in detecting hate speech.

4.4. Performance Comparison Before and After Class Imbalance Handling

Table 4 compares the performance of three methods—SVM, Naive Bayes, and IndoBERT—before and after handling class imbalance. After handling the imbalance, SVM showed improvements in all metrics: Accuracy (+14%), F1-Score (+21%), Recall (+26%), and Precision (+14%). Naive Bayes experienced a decrease in Accuracy (-5%) but significant increases in F1-Score (+14%), Recall (+16%), and Precision (+11%). IndoBERT also saw improvements in all metrics: Accuracy (+2%), F1-Score (+7%), Recall (+8%), and Precision (+5%). Overall, handling class imbalance improved the performance of all three methods, especially in terms of F1-Score, Recall, and Precision. Although SVM required longer training time, the performance improvement made it a strong choice.

5. Conclusions

This study evaluated the performance of SVM, Naive Bayes, and IndoBERT for hate speech detection in Indonesian using a multi-label dataset. The results indicate that IndoBERT outperforms both SVM and Naive Bayes, achieving an accuracy of 93% and consistently

high F1-score, recall, and precision at 91%. Effective preprocessing, including text normalization, stopword removal, and the use of TF-IDF for feature extraction improved model performance, particularly for SVM and Naive Bayes. Hyperparameter tuning and BERT-based data augmentation significantly boosted performance across all models, with IndoBERT showing a 30% accuracy improvement post-tuning. The findings underscore the importance of fine-tuning pre-trained models for language-specific tasks like hate speech detection, highlighting IndoBERT's superiority in handling complex linguistic contexts. The study contributes to the existing body of research by directly comparing these methods, revealing IndoBERT's potential for further improving hate speech detection in Indonesian and other low-resource languages. Future work could explore advanced data augmentation techniques and fine-tuning additional pre-trained models for broader applicability in real-time social media monitoring systems.

Author Contributions: Conceptualization, R.B., A.M., and Z.Z.; methodology, A.M. and R.R.; software, R.B., L.F. and S.A.N.; validation, A.M., Z.Z., and R.R.; formal analysis, R.B. and L.F.; investigation, R.B. and S.A.N.; resources, A.M. and S.A.N.; data curation, Z.Z., R.R. and L.F.; writing—original draft preparation, R.B., L.F. and S.A.N.; writing—review and editing, A.M., Z.Z. and R.R.; visualization, R.B.; supervision, A.M. and Z.Z.; project administration, A.M.; funding acquisition, A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This study does not receive external funding.

Ethical Clearance: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this study is available upon request from the corresponding author.

Conflicts of Interest: All the authors declare no conflicts of interest.

References

1. Azzaakiyyah, H. K. (2023). The Impact of Social Media Use on Social Interaction in Contemporary Society, *Technology and Society Perspectives (TACIT)*, Vol. 1, No. 1, 1–9. doi:10.61100/tacit.v1i1.33.
2. Castaño-Pulgarín, S. A., Suárez-Betancur, N., Vega, L. M. T., and López, H. M. H. (2021). Internet, Social Media and Online Hate Speech. Systematic Review, *Aggression and Violent Behavior*, Vol. 58, 101608. doi:10.1016/j.avb.2021.101608.
3. Bromell, D. (2022). Challenges in Regulating Online Content, *Regulating Free Speech in a Digital Age*, Springer International Publishing, Cham, 29–53. doi:10.1007/978-3-030-95550-2_2.
4. Toktarova, A., Syrlybay, D., Myrzakhmetova, B., Anuarbekova, G., Rakhimbayeva, G., Zhylanbaeva, B., Suieuova, N., and Kerimbekov, M. (2023). Hate Speech Detection in Social Networks using Machine Learning and Deep Learning Methods, *International Journal of Advanced Computer Science and Applications*, Vol. 14, No. 5. doi:10.14569/IJACSA.2023.0140542.
5. Rahman, A., Hidayati, I., Wahyuni, R., Kurniawan, R., and Putri, R. N. (2024). Problematic Internet Use at Junior and High School in Padang, Indonesia: The Interplay of Self-Esteem and Social Acceptance, *Participatory Educational Research*, Vol. 11, No. 5, 244–257. doi:10.17275/per.24.73.11.5.
6. Dianita, K. V. (2021). The Freedom of Speech Based on Jerinx Case, ITE Law Approach, *Journal of Digital Law and Policy*, Vol. 1, No. 1, 29–36. doi:10.58982/jdlp.v1i1.91.
7. Ayo, F. E., Folorunso, O., Ibharalu, F. T., and Osinuga, I. A. (2020). Machine Learning Techniques for Hate Speech Classification of Twitter Data: State-of-the-Art, Future Challenges and Research Directions, *Computer Science Review*, Vol. 38, 100311. doi:10.1016/j.cosrev.2020.100311.
8. Mullah, N. S., and Zainon, W. M. N. W. (2021). Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review, *IEEE Access*, Vol. 9, 88364–88376. doi:10.1109/ACCESS.2021.3089515.
9. Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., and Bahar, S. (2020). IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding, *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*, 843–857.
10. Shi, L., Liu, D., Liu, G., and Meng, K. (2020). AUG-BERT: An Efficient Data Augmentation Algorithm for Text Classification, 2191–2198. doi:10.1007/978-981-13-9409-6_266.
11. Abidin, Z., Junaidi, A., and Wamiliana. (2024). Text Stemming and Lemmatization of Regional Languages in Indonesia: A Systematic Literature Review, *Journal of Information Systems Engineering and Business Intelligence*, Vol. 10, No. 2, 217–231. doi:10.20473/jisebi.10.2.217-231.
12. Chai, C. P. (2023). Comparison of Text Preprocessing Methods, *Natural Language Engineering*, Vol. 29, No. 3, 509–553. doi:10.1017/S1351324922000213.
13. Ibrohim, M. O., and Budi, I. (2019). Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter, *Proceedings of the Third Workshop on Abusive Language Online*, Association for Computational Linguistics, Stroudsburg, PA, USA, 46–57. doi:10.18653/v1/W19-3506.
14. Wenando, F. A., and Fuad, E. (2019). Detection of Hate Speech in Indonesian Language on Twitter Using Machine Learning Algorithm, *Prosiding CELSciTech*, Vol. 4, 6–8.
15. Nugroho, K. S., Sukmadewa, A. Y., Wuswilahaken DW, H., Bachtiar, F. A., and Yudistira, N. (2021). BERT Fine-Tuning for Sentiment Analysis on Indonesian Mobile Apps Reviews, *6th International Conference on Sustainable Information Engineering and Technology 2021*, ACM, New York, NY, USA, 258–264. doi:10.1145/3479645.3479679.
16. Dharmawan, S., Mawardi, V. C., and Perdana, N. J. (2023). Klasifikasi Ujaran Kebencian Menggunakan Metode FeedForward Neural Network (IndoBERT), *Jurnal Ilmu Komputer Dan Sistem Informasi*, Vol. 11, No. 1. doi:10.24912/jiksi.v11i1.24066.
17. Yazid, R. M., Umbara, F. R., and Sabrina, P. N. (2023). Deteksi Ujaran Kebencian dengan Metode Klasifikasi Naïve Bayes dan Metode N-Gram pada Dataset Multi-Label Twitter Berbahasa Indonesia, *Informatics and Digital Expert (INDEX)*, Vol. 4, No. 2, 46–52. doi:10.36423/index.v4i2.894.
18. Sanya, A. D., and Suadaa, L. H. (2022). Handling Imbalanced Dataset on Hate Speech Detection in Indonesian Online News Comments, *2022 10th International Conference on Information and Communication Technology (ICoICT)*, IEEE, 380–385. doi:10.1109/ICoICT55009.2022.9914883.
19. Koto, F., Rahimi, A., Lau, J. H., and Baldwin, T. (2020). IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP, *ArXiv Preprint ArXiv:2011.00677*.
20. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention Is

All You Need, *Advances in Neural Information Processing Systems*.

21. Pen, H., Teo, N., and Wang, Z. (2024). Comparative Analysis of Hate Speech Detection: Traditional vs. Deep Learning Approaches, *2024 IEEE Conference on Artificial Intelligence (CAI)*, IEEE, 332–337. doi:10.1109/CAI59869.2024.00070.
22. Uysal, A. K., and Gunal, S. (2014). The Impact of Preprocessing on Text Classification, *Information Processing & Management*, Vol. 50, No. 1, 104–112. doi:10.1016/j.ipm.2013.08.006.
23. Yusliani, N., Primartha, R., and Diana, M. (2019). Multiprocessing Stemming: A Case Study of Indonesian Stemming, *International Journal of Computer Applications*, Vol. 182, No. 40, 15–19. doi:10.5120/ijca2019918476.
24. Nabiilah, G. Z., Alam, I. N., Purwanto, E. S., and Hidayat, M. F. (2024). Indonesian Multilabel Classification Using IndoBERT Embedding and Mbert Classification, *International Journal of Electrical & Computer Engineering (2088-8708)*, Vol. 14, No. 1.
25. Computer, J. H., Honova, S. M., Computer, V. P., Setiawan, C. A., Parmonangan, I. H., and Diana. (2023). Sentiment Analysis of Skincare Product Reviews in Indonesian Language using IndoBERT and LSTM, *2023 IEEE 9th Information Technology International Seminar (ITIS)*, IEEE, 1–6. doi:10.1109/ITIS59651.2023.10420222.
26. Noviandy, T. R., Idroes, G. M., Tallei, T. E., Handayani, D., and Idroes, R. (2024). QSAR Modeling for Predicting Beta-Secretase 1 Inhibitory Activity in Alzheimer's Disease with Support Vector Regression, *Malacca Pharmaceutics*, Vol. 2, No. 2, 79–85. doi:10.60084/mp.v2i2.226.
27. Noviandy, T. R., Idroes, G. M., Hardi, I., Afjal, M., and Ray, S. (2024). A Model-Agnostic Interpretability Approach to Predicting Customer Churn in the Telecommunications Industry, *Infolitika Journal of Data Science*, Vol. 2, No. 1, 34–44. doi:10.60084/ijds.v2i1.199.
28. Xu, Y., Hu, L., Zhao, J., Qiu, Z., Ye, Y., and Gu, H. (2024). A Survey on Multilingual Large Language Models: Corpora, Alignment, and Bias, *ArXiv Preprint ArXiv:2404.00929*.
29. Ferrer, L. (2022). Analysis and Comparison of Classification Metrics, *ArXiv Preprint ArXiv:2209.05355*.