The Performance Analysis of Hate Speech in Cyberbullying Cases Based on IndoBERT and Cendol

Nancy Olivia Syahanifa[1] Kartika Dwi Maharani[2] Anggara Budiyanto[3] Miftah Huljannah[4] Suryo Adhi Wibowo[5] Koredianto Usman[6]

[1-6]*Telecommunication Engineering Program Study, School of Electrical Engineering, Telkom University*

[*] *Corresponding author: korediantousman@telkomuniversity.ac.id*

Manuscript received June 19, 2025; revised July 17, 2025; accepted August 18, 2025

Abstract

The proliferation of hate speech on Indonesian social media, particularly during the 2024 general election period, has underscored the need for effective automated moderation mechanisms. This study compares the performance of two Indonesian language models, IndoBERT and Cendol, for hate speech detection on the X platform (formerly Twitter). A dataset of 8,375 comments was collected through keyword-based crawling and manually annotated into three categories: neutral, negative, and positive. Following preprocessing and oversampling to address class imbalance, both models were fine-tuned and evaluated. The findings indicate that IndoBERT performs more effectively in capturing formal linguistic patterns, whereas Cendol demonstrates advantages in recognizing informal and sarcastic expressions. These results suggest that each model exhibits distinct strengths depending on language characteristics and contextual usage. The study contributes insights toward the development of more context-aware AI-based moderation systems for the Indonesian digital environment.

Keywords: hate speech; NLP; IndoBERT; Cendol

DOI: 10.25124/jmecs.v12i2.9116

1. Introduction

The development of digital technology and social media has drastically transformed the landscape of communication in Indonesia. However, behind the ease of access and connectivity lies a serious issue: the rise in cyberbullying cases, particularly through hate speech. These types of comments often target individuals based on their ethnic background, religion, gender, or physical condition, and have a significant impact on victims' mental health [1]. In this context, social media platforms such as X (formerly Twitter) have become dominant spaces for the spread of hate speech, especially during crucial moments such as the 2024 general election [2], [3].

This phenomenon is not merely incidental but has become a systemic social issue. Several viral cases in the media underscore the urgency of this matter, such as the bullying of an elementary school student with special needs whose bruised condition was circulated online [4] and hate speech cases involving public figures that were taken to court [5]. Data from the Ministry of Communication and Information shows a 40% increase in cyberbullying reports from 2022 to 2024 [6]. These facts highlight the urgent need for an automated and efficient hate speech detection system to cope with the massive flow of data [7].

One of the main challenges in building such a system is the complexity of the Indonesian language as used on social media. Users tend to employ informal language, slang, abbreviations, and regional language mixes, which are difficult for rule-based conventional systems to process.

Modern Natural Language Processing (NLP) models like BERT have been widely adopted, but their limitations in understanding local context remain a problem. IndoBERT, a BERT-based language model trained on Indonesian corpora, achieved an F1-score of 76.32% in abusive sentence detection tasks [8], but its performance declined when faced with imbalanced data distribution wit F1-score 68.4% [9].

Meanwhile, other approaches such as BiLSTM, Cendol, Bert and SVM achieved higher accuracy, 82.29% to 87% respectively, yet still failed to grasp the semantic variations in informal language contexts [10], [11]. Recent advancements in Natural Language Processing (NLP) have significantly improved machine un-

derstanding of human language. Among these, the Transformer architecture has become the backbone of many state-of-the-art models, including BERT (Bidirectional Encoder Representations from Transformers) [12]. This architecture enables models to capture contextual relationships between words more effectively than traditional sequential models.

In the Indonesian context, IndoBERT is a language model trained on formal text corpora, while Cendol is optimized for informal, user-generated content. These models represent two contrasting yet complementary approaches to handling linguistic diversity on social media, making them suitable candidates for hate speech detection tasks. To address these issues, this study aims to analyze and compare the performance of two Indonesian NLP models: IndoBERT and Cendol.

IndoBERT excels at processing formal and explicit text, while Cendol is designed to handle local and informal expressions commonly used on social media. Each model presents distinct advantages and limitations: IndoBERT performs well in recognizing structured and policy-related language but tends to struggle with slang and informal content. Conversely, Cendol is more effective in interpreting casual, modified, and sarcastic comments, but may underperform in formal contexts. These complementary characteristics suggest the potential benefit of ensemble learning in future work.

This study focuses solely on evaluating both models independently to assess their respective strengths and weaknesses in hate speech detection. This research involves the following stages: (1) collecting 8,375 tweet data from the X platform, (2) preprocessing data including text cleaning and normalization, (3) training the models using various hyperparameter settings, and (4) evaluating model performance using metrics such as accuracy, precision, recall, and F1-score [13]. Challenges such as class imbalance, ambiguity in slang, and computational limitations of large models are addressed through strategies like fine-tuning, oversampling, data augmentation, and batch size adjustments [14], [15] [16], [17].

2. Research Method

2.1. Data Collection

The dataset was collected from platform X (formerly Twitter) using a keyword-based crawling method. A list of 64 hate-related keywords was generated through an open survey involving 328 respondents. The open survey involved 328 Indonesian respondents aged 17 to 35 years, primarily consisting of university students, early-career professionals, and active social media users. Participants were recruited through online sharing on academic and community platforms. While the sample represents a digitally literate segment of the population, it may reflect some sampling bias toward younger and more urban-based demographics. Nevertheless, this group aligns with the primary audience and user base of social media platforms where hate speech is commonly found. These keywords included commonly used derogatory terms such as "`anjing`", "`bangsat`", and "`bodoh`". The crawling process yielded 8,375 raw comments written in Indonesian. The dataset focused on public posts containing potential elements of cyberbullying or hate speech, particularly during politically sensitive periods.

2.2. Data Processing

The collected comments underwent several preprocessing steps to ensure data quality and input consistency. These included text cleaning (removal of mentions, hashtags, URLs, emojis, and special characters), normalization of informal expressions using a custom dictionary, tokenization, stopword removal, and stemming. While common function words such as dan, yang, and atau were removed, important negations like tidak and bukan were deliberately retained to preserve sentence meaning. Although Transformer-based models such as IndoBERT and Cendol can process raw text directly, these preprocessing steps were applied conservatively to reduce noise, particularly in informal and user-generated content. Preliminary experiments showed that this approach improved training stability without significantly compromising contextual understanding.

2.3. Data Labeling Process

Each comment was manually annotated into one of three classes:

- Neutral (0)
- Negative (1)
- Positive (2)

The labeling process was conducted manually by the researcher based on defined guidelines for each class. Although only a single annotator was involved, careful attention was paid to consistency by referring to labeling criteria and conducting multiple passes to cross-check ambiguous cases. Each comment was contextually reviewed to determine whether it fell under the neutral, negative (hate speech) or positive (counter speech) category. To reduce subjectivity, the labeling was guided by commonly accepted definitions of hate speech in Indonesian digital discourse, including keywords and contextual tone. The classification used in this study adopts a three-class labeling scheme specifically tailored for the context of hate speech detection. Rather than representing emotional sentiment, these labels reflect the communicative intent of the comment in relation to hate speech. Neutral comments are those that do not contain harmful, offensive, or emotionally charged language and are generally factual, irrelevant, or non-engaged in nature. Negative comments explicitly or implicitly contain hate speech, including offensive expressions, personal insults, or derogatory remarks aimed at individuals or groups. In contrast, Positive comments represent counter speech—comments that aim to defend victims, promote tolerance, or reduce hostility by responding to hate in a constructive manner.

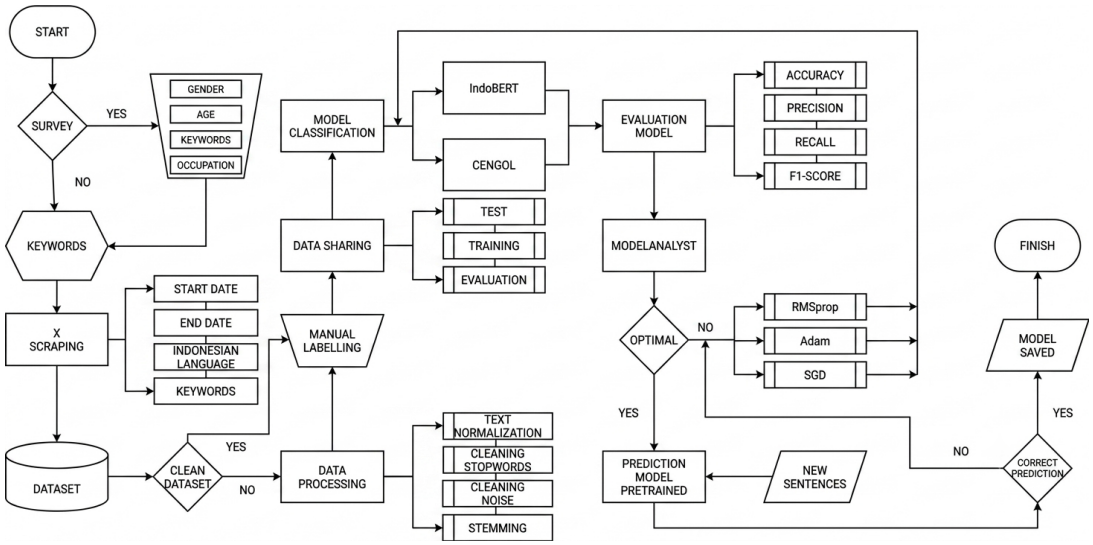


Fig. 1. Methodological flowchart for IndoBERT and Cendol evaluation

2.4. Class Balancing Strategy (SMOTE)

To address the imbalance in the original dataset, the Synthetic Minority Over-sampling Technique (SMOTE) was applied before splitting the data into training and testing sets.

SMOTE was implemented on the numerical feature vectors generated via TF-IDF representation rather than on the raw text [18]. To prevent data leakage, the dataset was split prior to oversampling, ensuring that synthetic sample generation was strictly restricted to the training set. Through interpolation between existing minority class feature vectors, SMOTE balanced the dataset across all three classes (negative, neutral, and positive), resulting in an equal distribution with a total of 8,375 instances.

2.5. Data Training and Evaluation

After applying SMOTE to balance the dataset, we randomly split the data into 90% for training and 10% for testing. Since the dataset was already balanced across all classes (neutral, negative, positive), the class distribution remained consistent in both sets, eliminating the need for stratified sampling

- Epochs: 5
- Batch size: 2 to 8
- Learning rate: 10^{-5} to 10^{-4}
- Sequence length: 256 tokens

The models were trained and evaluated using the following metrics:

- Accuracy
- Precision
- Recall
- F1-score

Training and validation performance were tracked over epochs using loss and accuracy plots to monitor overfitting. Confusion matrices were used to analyze classification behavior per class.

2.6. Technical Overview of IndoBERT and Cendol

IndoBERT is a Transformer-based model derived from BERT and pretrained on over 5.5 billion words of Indonesian formal texts such as Wikipedia, news, and government documents. With approximately 125 million parameters, IndoBERT is optimized for understanding structured language, formal expressions, and socio-political contexts. It uses bidirectional attention to capture word context from both left and right directions, which improves understanding of sentence semantics in formal registers. Cendol, on the other hand, is a larger model with around 580 million parameters, specifically trained on Indonesian social media text. Its tokenizer and vocabulary are tailored to handle slang, abbreviations, spelling variations (e.g., "`b*doh`" or "`anj1ng`"), and implicit expressions often found in online discourse. Cendol excels in detecting sarcasm, disguised hate speech, and informal tone, which are limitations in traditional models or models trained on formal corpora.

The architectural and corpus variations between the two models yield complementary performance. Figure 1 illustrates the comprehensive experimental framework of this study, encompassing data collection, preprocessing, model training, evaluation, and optimization for both IndoBERT and Cendol. IndoBERT is more reliable for analyzing formal hate speech such as in news or legal discussions, while Cendol performs better in identifying casual, veiled, or coded hate speech in social platforms.

Before the data was used for training the NLP models, several preprocessing steps were carried out to ensure optimal data quality. The first step was data cleaning, which involved removing irrelevant elements

from the text such as mentions (`@username`), hashtags (#), links (URLs), and non-alphabet characters. For example, the sentence "`@user123 ini bodoh banget sih` (@user123 is so dumb)" was transformed into "`ini bodoh banget sih` (it's so dumb)". The next step was text normalization, which involved converting informal or non-standard words into their standard forms according to Indonesian language rules. This process used a custom dictionary developed based on popular word variations found on social media. For example, "`gpp` (it's okay (informal words))" was normalized to "`tidak apa-apa` (it's okay (formal words))", and "`lo` (you (informal word))" to "`kamu` (you (formal word))" (you).

After normalization, stopwords removal was performed. Stopwords are high-frequency words with low semantic contribution, such as "dan" (and), "atau" (or), and "dengan" (with). The stopword list was compiled from Indonesian NLP resources and matched against sentence structures in the dataset. The next step was stemming, which converts words to their root forms to preserve semantic consistency in text feature representation. This process utilized the Sastrawi library. For example, the word "`membenci`" (to hate) was stemmed to "`benci`" (hate).

Overall, the data collection and processing stages in this study were designed to produce a representative, clean, and balanced dataset. This process is a crucial foundation for training reliable and accurate NLP models, especially in the context of the Indonesian language, which is characterized by high linguistic diversity. With this approach, the model is expected to recognize hate speech in various forms of expression —both explicit and implicit— and in both formal and informal language.

3. Result and Discussion

3.1. Overview of Data

The dataset used in this study was obtained from the social media platform X (formerly Twitter) through a keyword-based crawling technique. The data collection process focused on public comments that potentially contain elements of cyberbullying and hate speech. The keywords used for crawling were determined through an open survey involving 328 respondents. From this survey, 64 keywords commonly associated with hate speech were identified, such as "`anjing`" (dog), "`bangsat`" (bastard), "`bodoh`" (stupid), among others. These words represent verbal expressions commonly used in online bullying practices in Indonesia and served as the basis for the data collection script. The crawling process yielded a total of 8,375 comments as the initial dataset.

The label distribution of the original dataset showed a significant class imbalance, with negative comments dominating at 2,150 samples or 55.5%, followed by neutral comments with 1,200 samples (31%), and positive comments with only 525 samples (13.5%). This imbalance has the potential to reduce the performance of classification models due to bias toward the majority class. Therefore, the Synthetic Minority Oversampling Technique (SMOTE) was applied to balance the dataset.

After oversampling, the total number of data samples increased to 8,375, with a more proportional label distribution: 3,500 negative (41.8%), 3,000 neutral (35.8%), and 1,875 positive (22.4%).

3.2. IndoBERT Model Performance

The IndoBERT model achieved optimal performance at the 5-th epoch with the following hyperparameter configuration: a learning rate of 10^{-5}, batch size of 8, and sequence length of 256. Evaluation was conducted using 10% of the total dataset as test data. The evaluation results showed an overall accuracy of 90.7%, with an average precision of 88.9%, recall of 89.0%, and F1-score of 88.9%. The detailed classification performance of the model for each label is presented in Table 1. For the neutral label, the model recorded a precision of 91%, recall of 86%, and F1-score of 88%. For the negative label, recall was relatively high at 93%, although its precision was lower compared to other classes (86%). Meanwhile, the positive label achieved the highest precision at 99% and an F1-score of 97%, indicating high effectiveness in recognizing positively toned comments, despite the limited number of samples in this class.

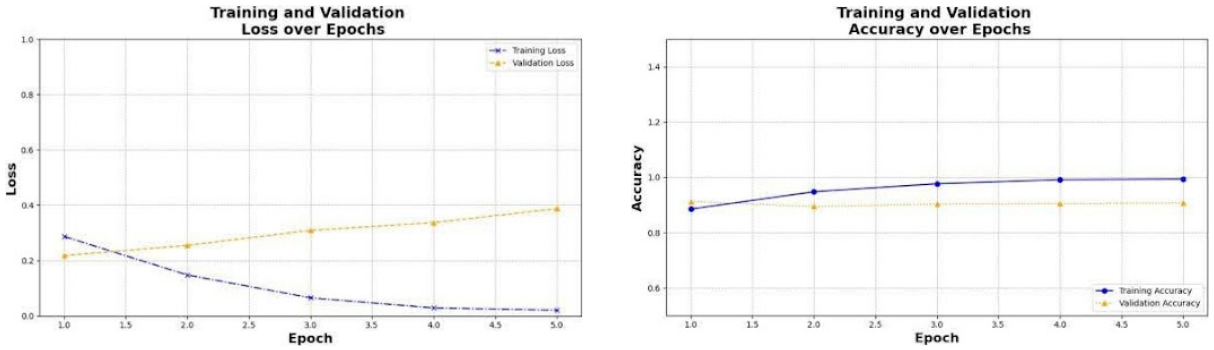
Table 1: IndoBERT Performance Metrics by Class

Class	Acc.	Prec.	Rec.	F1-Score
Neutral	92%	91%	86%	88%
Negative	92%	86%	93%	89%
Positive	98%	99%	95%	97%

The training trend of the model is visualized in two main graphs. Figure 2(a) shows a consistent decrease in training loss from 1.52 to 0.21, while the validation loss fluctuates slightly and tends to increase in the later epochs. This divergence between training and validation loss may indicate early signs of overfitting. It suggests that although the model effectively learns patterns from the training data, its ability to generalize to unseen data might have started to decline. Future improvements could include applying early stopping, dropout, or additional regularization to mitigate this issue [15, 20].

Next, Figure 2(b) provides a clear and detailed illustration of the progressive increase in training accuracy, which rises significantly from an initial value of 85.6% to a much higher level of 95.2% over the span of five training epochs. Alongside this, the validation accuracy also shows a meaningful improvement, increasing from 82.3% to 90.1% during the same period.

This steady upward trend in both training and validation accuracies is an important indicator of the model's successful learning process.

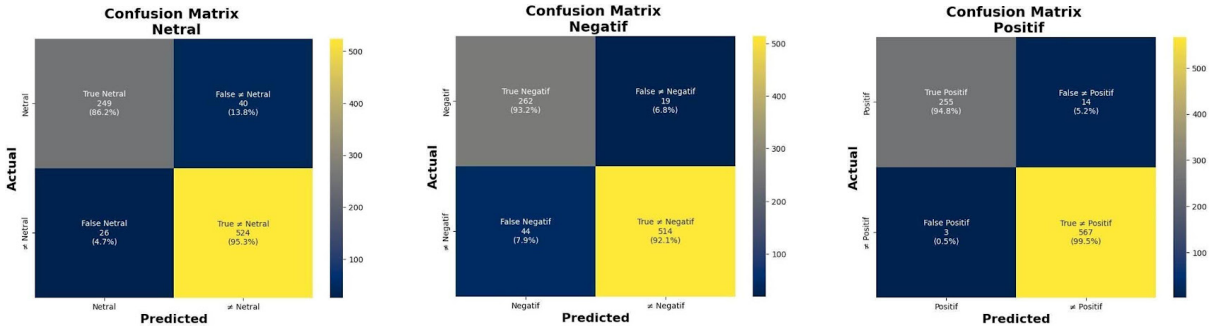


(a) (b)

Fig. 2. Model performance evaluation of IndoBERT: (a) Training versus Validation Loss, and (b) Training versus Validation Accuracy

.

G



(a) (b) (c)

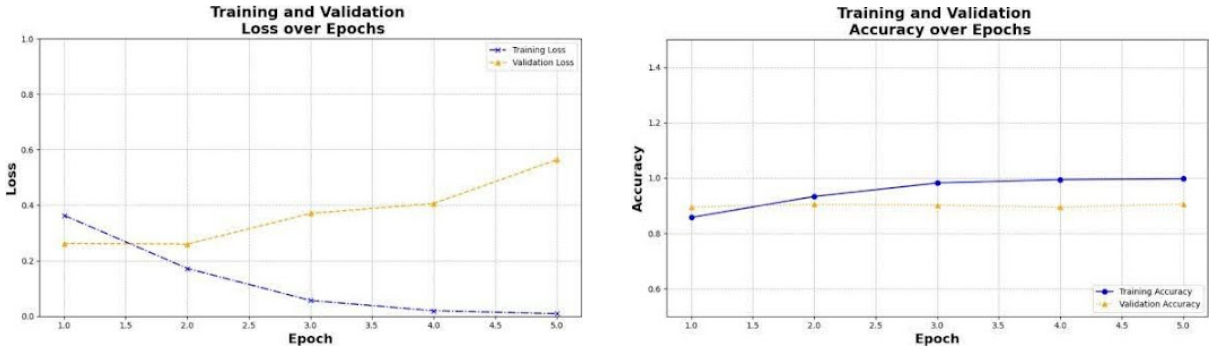
Fig. 3. Comparison of model performance using Confusion Matrix across different cases: (a) Neutral, (b) Negatif, and (c) Positive, for IndoBERT model.

It suggests that the training regimen is effectively guiding the optimization of the model parameters, enabling the model to better fit the training data while simultaneously improving its ability to generalize to new, unseen data. Such improvements are crucial, as they imply that the model does not simply memorize the training examples but is genuinely learning meaningful patterns that can be applied more broadly. Overall, these results underscore the effectiveness of the training strategy and highlight the model's growing predictive performance and robustness as it undergoes further training.

To strengthen the quantitative analysis, Figure 3(a), 3(b) and 3(c) present the confusion matrix for each label. This visualization shows true positives and misclassifications for neutral, negative, and positive labels. The correct classification rates are 89.2% for neutral, 92.1% for negative, and 95.4% for positive labels. However, the model still exhibits difficulty distinguishing between neutral and negative comments, which often have high semantic ambiguity.

The main advantage of IndoBERT lies in its pretraining process using a very large Indonesian language corpus, approximately 5.5 billion words, covering online news, Wikipedia, legal documents, and discussion forums. This enables the model to understand formal vocabulary and socio-political contexts, such as SARA (Ethnicity, Religion, Race, and Inter-group) terms [9], [19]. For example, phrases like "discriminatory policy" are classified as negative comments because they are recognized as expressions of injustice. Additionally, IndoBERT's bidirectional attention architecture allows it to capture contextual meaning from both directions. A relevant example is the sentence "He is not a stupid person, but unlucky," where the word "stupid" is not classified as negative because the defensive context is detected by the model [16].

Hyperparameter configuration also plays a crucial role in training stability. Using a low learning rate such as 10^{-5} and a small batch size such as 8 helps reduce gradient fluctuations and prevents model divergence during training [14], [20], [21].

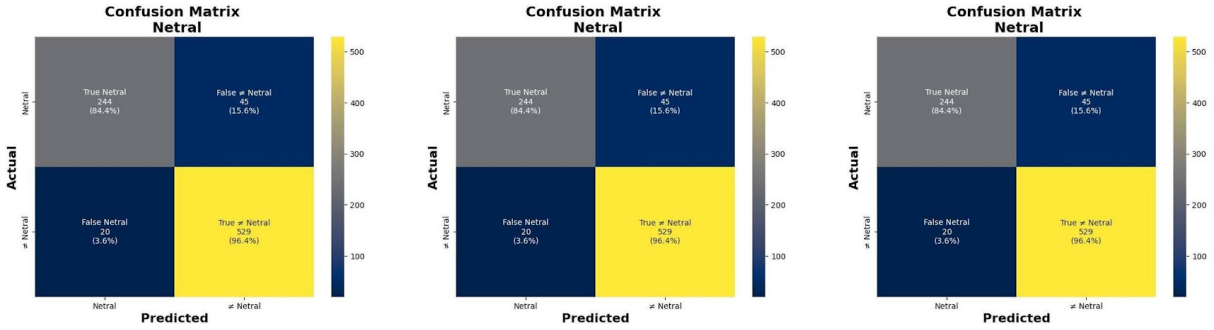


(a) (b)

Fig. 4. Model performance evaluation of Cendol: (a) Training versus Validation Loss, and (b) Training versus Validation Accuracy

.

G



(a) (b) (c)

Fig. 5. Comparison of model performance using Confusion Matrix across different cases: (a) Neutral, (b) Negatif, and (c) Positive, for Cendol model.

Linguistic ambiguity: Comments with sarcastic tones, such as "Amazing, all you do is complain," are often difficult to identify as negative since the literal form appears positive.

Previous research using IndoBERT for Indonesian-language hoax detection reported an accuracy of 91.2% and about similar value [9]. In contrast, LSTM-based models achieved only 82.3% accuracy on a similar task due to limitations in capturing word relationships in longer contexts. This confirms IndoBERT's superiority in understanding Indonesian syntactic structures, especially in formal and semi-formal texts.

3.3. Cendol Model Performance

The Cendol model is designed to capture the distinctive linguistic expressions commonly found in social media communication in Indonesia, particularly in comments containing hate speech. This model employs a transformer-based approach, which has been fine-tuned using an Indonesian social media corpus to better understand local informal language patterns [10]. In this study, the model was tested using optimal configuration, which includes a learning rate of 10^{-4}, batch size of 2, and five epochs. Quantitative evaluation results show that Cendol achieved an overall accuracy of 90.6%, with an average precision of 89.5%, recall of 88.9%, and F1-score of 89.2%. Its performance remained consistent and stable, even under imbalanced data conditions before class balancing was applied. Overall, these results demonstrate that the Cendol architecture is highly suitable for sentiment-based and hate speech classification tasks in Indonesian social media contexts [9], [11]. When performance is analyzed by individual labels, the model also shows very competitive results. Table 2 indicates that the neutral label achieved 92% precision and accuracy, although its recall was lower at 84%, suggesting that some neutral comments were misclassified. Meanwhile, the negative label recorded the highest recall at 95%, with a moderate precision of 86%. The strongest performance was observed on the positive label, which achieved both 97% precision and 95% recall, resulting in an F1-score of 96%.

To assess the stability of the model's training process, Figure 4(a), 4(b), and 4(c) presents the graphs of training and validation loss, as well as training and validation accuracy across epochs. The loss graph shows a

Table 2: Cendol Performance Metrics by Class

Class	Acc.	Prec.	Rec.	F1-Score
Neutral	92%	92%	84%	88%
Negative	93%	86%	95%	90%
Positive	98%	97%	95%	96%

consistent decrease from 1.2 in the first epoch to 0.3 by the fifth epoch, while validation loss remains stable in the range of 0.4–0.5. This indicates that overfitting did not occur, as the gap between training and validation loss is not significant. Figure 5(a), 5(b) and 5(c) present the confusion matrix for each label.

Training accuracy increased from approximately 85% to 95% by the fifth epoch, while validation accuracy remained steady at around 90.6%. This trend suggests that the model has a strong generalization capability toward unseen test data, even under the diverse and dynamic conditions commonly found on social media platforms.

Similar to IndoBERT, Cendol's training loss decreased consistently, while the validation loss fluctuated and showed a slight upward trend in the final epochs, as seen in Figure 4. This may also indicate a tendency toward overfitting, suggesting the need for further regularization techniques in future work. Figure 4 further reinforces the quantitative findings by presenting a detailed visualization of the confusion matrix encompassing all three classification labels. This graphical representation allows for a clearer understanding of how well the model performs in distinguishing between the different classes. The model demonstrates consistently high classification accuracy across each of the categories, with a true positive rate of 88.4% for the neutral label, indicating strong performance in correctly identifying neutral instances. Additionally, the model achieves an impressive true positive rate of 95.1% for the negative label, reflecting its effectiveness in accurately detecting negative cases. Furthermore, the highest true positive rate of 96.1% is observed for the positive label, showcasing the model's robust capability to correctly classify positive instances. Together, these results highlight the model's overall reliability and precision in classification tasks, as evidenced by the well-distributed and high accuracy scores depicted in the confusion matrix visualization. However, a small number of misclassifications can still be observed, such as neutral comments being classified as negative, which is common in social text with ambiguous meanings.

Figure 5 displays the confusion matrices for the Cendol model, highlighting its ability to classify neutral, negative, and positive comments with high accuracy across all categories. Cendol is specifically designed to handle the informal linguistic expressions frequently used on Indonesian social media platforms such as X (formerly Twitter), TikTok, and Instagram. The model leverages a tokenizer that is more adaptive to variations in text, including abbreviations, slang, and alternative spellings, such as "anjing (dog)" → "4nj1ng", or "bodoh (dumb)" → "b*doh". This capability stems from Cendol being trained on a corpus rich in Indonesian netizen speech styles, in contrast to more formal models like IndoBERT. Cendol's effectiveness in detecting hate speech is also evident in its ability to capture implicit meaning in sarcastic comments, such as the phrase "lu emang pinter banget ya" ("oh you're really so smart, huh"), which is identified as a negative comment based on the semantic contrast within the sentence. This shows that Cendol does not rely solely on pattern matching of offensive words but also understands inter-word relationships in a sentence—an advantage that rule-based or classical models like SVM do not possess [10], [22].

Furthermore, the use of this model is highly relevant to the dynamic nature of digital communication in Indonesia, characterized by evolving slang and camouflage strategies (e.g., the use of numbers or symbols to evade platform moderation systems). In such contexts, a model like Cendol is essential for platform security developers aiming to automatically detect harmful content. The findings of this study indicate that locally trained models outperform generic models in the classification of hate speech. In previous research, approaches using Cendol even achieved over 90% accuracy when implemented in Cybersentinel applications based on VADER and grid search optimization [23]. In contrast, traditional methods like Naive Bayes or Logistic Regression only achieved accuracies of 72–85%, highlighting the superiority of local deep learning models like Cendol [24]. These results align with previous findings [9], [19], [25], [3], [16], [21], which suggest that BERT-based models such as IndoBERT tend to exhibit high stability in classifying formal text. IndoBERT is more reliable when classifying comments written in standard or formal language structures, whereas Cendol demonstrates stronger performance when handling informal language and sarcastic expressions.

A visual analysis of the training curves reveals key differences in the dynamics of both models. IndoBERT converges more quickly, reaching peak validation accuracy at the third epoch. In contrast, Cendol requires five epochs to reach its maximum accuracy. Additionally, IndoBERT achieves a lower validation loss (approximately 0.35) compared to Cendol (around 0.45), indicating better generalization performance on formal data.

These differences are further confirmed through the confusion matrix analysis [13]. IndoBERT shows lower misclassification rates for neutral and negative labels, while Cendol performs more accurately in classifying positive comments. This has important implications for selecting the appropriate model depending on the dominant type of user expression. Factors Affecting Performance Differences:

3.3.1. Formal vs. Informal Context

IndoBERT was trained on a corpus of 5.5 billion words sourced from formal domains such as Wikipedia,

Table 3: Comparison between IndoBERT and Cendol Models

Aspect	**IndoBERT**	**Cendol**
Accuracy	90.7%	90.6%
Precision	88.9%	89.5%
Recall	89.0%	88.9%
F1-Score	88.9%	89.2%
Strengths	Strong on formal language, legal, policy-related, and structured content	Excels in informal text, slang, sarcasm, disguised and ambiguous expressions
Weakness	Less effective in detecting informal, sarcastic, or code-mixed expressions	Less consistent in handling formal or technical vocabulary
Number of Parameters	Approximately 125 million	Approximately 580 million
Training Corpus	Formal Indonesian	Informal Indonesian

Table 4: Error Analysis and Misclassification

Examples Text & Comment	**True Label**	**IndoBERT**	**Cendol**	**Remarks**
"Wah kamu pinter banget sih"	Negative	Positive	Negative	Sarcasm. IndoBERT failed to detect sarcastic tone, Cendol succeeded.
"Kebijakan ini benar-benar merugikan rakyat"	Negative	Neutral	Negative	IndoBERT misinterpreted formal language as neutral.
"Hadeh, ini tuh lucu banget sumpahhh"	Neutral	Positive	Positive	Both models incorrectly picked up emotional intensity as positivity.
"Bagus banget, ga seperti biasanya. Hebat deh."	Neutral	Positive	Positive	Sarcastic praise is misclassified by both models.

news articles, and legal documents [16]. As a result, the model excels at understanding technical terminology and structured language. For example, phrases like "violation of Article 28 of the ITE Law" are easily recognized by IndoBERT as hate speech with a legal orientation. On the other hand, Cendol was developed using data from social media platforms rich in slang, informal expressions, and symbol substitutions (e.g., "anj1ng", "b4ngs4t"). This allows Cendol to effectively detect disguised or veiled hate speech, including code-mixed comments that are common on platforms like X (formerly Twitter) and TikTok.

3.3.2. Model Size and Parameter Complexity

IndoBERT contains approximately 125 million parameters, making it a lightweight and resource-efficient model [9]. In contrast, Cendol contains about 580 million parameters, giving it greater capacity to capture complex semantics but also requiring more computational resources. The larger parameter size of Cendol allows it to be more flexible in processing non-standard language structures. However, it also necessitates more sensitive fine-tuning, including careful hyperparameter adjustments and dropout optimization strategies.

3.3.3. Sensitivity to Class Imbalance

comments is class imbalance. In this study, oversampling was applied using SMOTE to balance the class distribution. IndoBERT demonstrates relatively higher resilience to skewed data distributions, whereas Cendol requires more intensive data augmentation strategies. The comparison between the two models demonstrates that model selection should be tailored to the specific context of use. IndoBERT is more suitable for systems that analyze legal content, public policy, and political disinformation, where formal language and structured narratives are commonly used . In contrast, Cendol is more effective for detecting comments on social media, making it particularly useful for anti-cyberbullying campaigns and online community moderation.

Therefore, the optimal approach is to integrate both models within an ensemble learning or voting system a strategy also supported by recent NLP studies. Such a system can combine IndoBERT's strength in understanding formal and legal language with Cendol's ability to capture the informal and dynamic expressions typical of online communication.

3.3.4. Comparative Analysis and Error Examination

To gain a deeper understanding of each model's capabilities, a comparative evaluation between IndoBERT and Cendol was conducted. As shown in Table 3. Both models achieved similar overall performance, with IndoBERT slightly leading in formal text classification and Cendol excelling in informal and sarcastic expressions. IndoBERT reached 90.7% accuracy and demonstrated robustness in processing structured con-

tent, while Cendol achieved 90.6% accuracy and was more effective in capturing disguised or casual hate speech.

In addition to quantitative performance, Table 4 presents an error analysis based on real examples. It reveals typical misclassification scenarios that highlight the models' respective limitations. IndoBERT often fails to detect sarcasm or disguised negativity, while Cendol, despite its informal fluency, sometimes misclassifies neutral content as positive due to emotional language.

The examples above indicate that both models struggle with sarcastic, ambiguous, or emotionally charged expressions. IndoBERT, while strong in formal language processing, tends to misclassify comments with disguised negative intent. Cendol performs better with informal and sarcastic tones but may still face difficulties with nuanced interpretation. These findings suggest that future improvements could include sarcasm-aware training strategies or attention-based postprocessing.

4. Conclusion

This study aims to compare the performance of two Indonesian language-based Natural Language Processing (NLP) models, IndoBERT and Cendol, in detecting hate speech on the social media platform X (Twitter). Based on quantitative evaluation results, both models demonstrate high accuracy—90.7% for IndoBERT and 90.6% for Cendol. IndoBERT proves superior in handling formal and structured texts, thanks to its training on a large Indonesian corpus sourced from official materials such as news articles, legal documents, and Wikipedia. In contrast, Cendol is specifically designed to understand informal language, including slang, symbols, and non-standard spelling commonly used on social media, enabling it to capture implicit meanings in sarcastic or coded expressions.

These differences highlight the importance of selecting models according to their intended use case. IndoBERT is more suitable for formal content analysis systems, such as public policy reporting or legal disinformation detection, while Cendol is ideal for moderating user comments on social media with casual language styles. To enhance the accuracy and adaptability of hate speech detection systems, this study recommends implementing an ensemble learning approach that synergistically combines the strengths of both models.

Suggestions for future research include developing multimodal approaches that analyses not only text but also visual elements like images and videos in memes or provocative content, allowing detection systems to better respond to the complexity of today's digital communication. Additionally, exploring domain-specific hate speech, such as gender, sexual orientation, and disability, could broaden the ethical and social contributions of this research.

Acknowledgment

This research was conducted with the support of the AI Center Telkom University – Center of Excellence Artificial Intelligence for Learning and Optimization (AILO).

Author Information

Nancy Olivia S study at the department of Electrical Engineering (Telecommunications), Telkom University, Bandung, Indonesia since 2020 for her bachelor's degree (S.T.) in 2024. Since September 2021 Nancy Olivia Syahanifa is with the school of electrical engineering, Telkom University, Bandung, Indonesia as a research assistant of The Center of Advanced Intelligent Communications (AICOMS) (formerly the Center for Advanced Wireless Technologies (AdWiTech) 2016-2020).

Kartika Dwi M. graduated from Bachelor Degree of Telecommunication Engineering, Faculty of Electrical Engineering, Telkom University. Her research interest include: signal processing, deep learning, and data analysis. She is highly interested in Artificial Intelligence (AI) and she completed a thesis focusing on the performance analysis of AI NLP models, such as IndoBERT and Cendol, for sentiment analysis in cyberbullying cases.

Anggara Budiyanto graduated from Telecommunications Engineering, Faculty of Electrical Engineering, Telkom University. He is highly interested in Artificial Intelligence (AI), particularly in Natural Language Processing (NLP). Completed a thesis focusing on the performance analysis of AI NLP models, such as IndoBERT and Cendol, for sentiment analysis in cyberbullying cases. Additionally, has a keen interest in Quantum Information, specifically the concept of negative entropy.

Miftah Huljannah graduated from Telecommunication Engineering Telkom University with expertise in Data Analytics, Artificial Intelligence, Machine Learning, Deep Learning, and Intelligent Automation. Experienced in transforming operational and network data into actionable insights through KPI

analysis, reporting automation, and AI-driven solutions. She is interested in Natural Language Processing (NLP) and transformer-based deep learning models, including IndoBERT for hate speech and cyberbullying detection on Indonesian social media datasets. Experienced in text preprocessing, tokenization, feature engineering, model evaluation, and machine learning pipeline development.

Suryo Adhi Wibowo graduated from Bachelor of Telecommunication Engineering and Master of Electrical Engineering at Faculty of Electrical Engineering, Telkom University. He finished his Ph.D from Pusan National University of Korea in 2018. He joined as lecturer in Telecommunication Engineering, Telkom University after he was graduated in 2012. He was the head researcher at Image Processing and Vision. He was also a researcher at AI for Learning and Optimization (AILO) research center. His main interest are in signal processing, AI and smart systems, as well as data analysis.

Koredianto Usman joined Telkom University (formerly known as STTTelkom) in 2002. He graduated his bachelor from electrical engineering ITB in 1999, and his master from telecommunication engineering TUM Germany in 2001. He graduated his doctorate from School of Electrical Engineering and Informatics ITB in 2019. His research interest includes: signal processing, compressive sensing and open sources system.

Author Contributions

Nancy Olivia Syahanifa: Conceptualization, methodology, investigation, data curation, and writing – original draft.
Kartika Dwi Maharani: Methodology, formal analysis, validation, and visualization.
Anggara Budiyanto: Software, investigation, formal analysis, and validation.
Miftah Huljannah: Data curation, visualization, and formal analysis.
Suryo Adhi Wibowo: Supervision, validation, and writing – review and editing.
Koredianto Usman: Supervision, conceptualization, validation, and writing – review and editing.

ORCID ID

Suryo Adhi Wibowo (https://orcid.org/0000-0002-3084-8534)
Koredianto Usman (https://orcid.org/0009-0003-7091-1774)

AI Usage Policy

Generative artificial intelligence tools were used only for language editing and LaTeX formatting assistance. All scientific content, analyses, results, and conclusions were reviewed and verified by the authors, who take full responsibility for the content of this manuscript.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Open Access Policy

References

[1] N. Agustiningsih, A. Yusuf, A. Ahsan, and Q. Fanani, "The impact of bullying and cyberbullying on mental health: a systematic review," *International Journal of Public Health Science*, vol. 13, no. 2, 2024.

[2] N. Muhamad, "Twitter, medsos dengan ujaran kebencian terbanyak pada kampanye pemilu 2024," Databoks Katadata, Feb. 2024, https://databoks.katadata.co.id/teknologi-telekomunikasi.

[3] A. B. Y. A. Putra and Y. Sibaroni, "Disinformation detection on 2024 Indonesia presidential election using IndoBERT," in *2023 International Conference on Data Science and Its Applications (ICoDSA)*, 2023, pp. 350–355.

[4] F. Nadhiroh, "Fakta siswi SD diviralkan korban perundungan ternyata berkebutuhan khusus," detikJatim, Oct. 2024, https://www.detik.com/jatim/berita/d-7592000.

[5] R. Naibaho, "Berkas perkara kasus ujaran kebencian soal Papua tiktoker AB lengkap," detikNews, Feb. 2024, https://news.detik.com/berita/d-7206759.

[6] L. Rizkinaswara, "Strategi kominfo cegah cyberbullying saat pembelajaran daring," Kementerian Kominfo, Sep. 2020, accessed: Nov. 9, 2024. https://aptika.kominfo.go.id/2020/09/strategi-kominfo.

[7] H. Putri, "Dosen UNM ciptakan aplikasi CyberSentinel untuk deteksi cyberbullying," Nusantara Post, Aug. 2024, https://nusantara-post.com/gaya-hidup/dosen-unm.

[8] A. R. Hanum, I. A. Zetha, S. C. Putri *et al.*, "Analisis kinerja algoritma klasifikasi teks bert dalam mendeteksi berita hoaks," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 3, pp. 537–546, 2024.

[9] H. Putra, M. A. Bijaksana, and A. Romadhony, "Deteksi penggunaan kalimat abusive pada teks bahasa indonesia menggunakan metode IndoBERT," *Jurnal Tugas Akhir Fakultas Informatika*, vol. 8, no. 2, pp. 3028–3038, 2021.

[10] F. A. Artanto, "Support vector machine berbasis particle swarm optimization pada analisis sentimen anggota KPPS," *Jurnal FASILKOM*, vol. 14, no. 1, pp. 75–79, 2024.

[11] "IndoNLP, Cendol, Hugging Face," 2024, accessed: 2024-10-18.

[12] A. Gillioz, J. Casas, E. Mugellini, and O. A. Khaled, "Overview of the transformer-based models for nlp tasks," in *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, Sep. 2020, pp. 179–183.

[13] Rina, "Memahami confusion matrix: Accuracy, precision, recall, specificity, dan f1-score untuk evaluasi model klasifikasi," https://esairina.medium.com/memahami-confusion-matrix, 2023, accessed: 16-10-2024.

[14] M. Cacic, "What are fine-tuning hyperparameters and how to set them," https://www.entrypointai.com/blog/fine-tuning-hyperparameters/, August 2023, accessed: 2024-11-10.

[15] L. Craig, "What is fine-tuning in machine learning and ai?" https://www.techtarget.com/searchenterpriseai/definition/fine-tuning, Jul. 2024, accessed: 2024-11-10.

[16] L. R. Aini, E. Nurfadhilah, A. Jarin, A. Santosa, and M. T. Uliniansyah, "Enhancing sentiment analysis models through multi-technique data augmentation: A study with indobert," in *2023 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, 2023, pp. 137–142.

[17] S. Fathima, "Learn how to do model fine-tuning: Challenges, metrics and best practices," https://www.markovml.com/blog/model-fine-tuning, Dec. 2023, accessed: 2023-12-05.

[18] R. Tracker, "Apa itu tf-idf?" https://www.ranktracker.com/id/seo/glossary/td-idf/, 2022, accessed: 2024-11-09.

[19] Rahmawati, A. Alamsyah, and A. Romadhony, "Hoax news detection analysis using IndoBERT deep learning methodology," in *2022 10th International Conference on Information and Communication Technology (ICoICT)*, 2022, pp. 368–373, https://api.semanticscholar.org/CorpusID:253047025.

[20] "Beyond basics: 5 fine-tuning stages for precision in machine learning," https://hyperight.com/beyond-basics, May 2023, accessed: 2024-11-10.

[21] P. Nabila and E. B. Setiawan, "Adam and adamw optimization algorithm application on bert model for hate speech detection on twitter," in *2024 International Conference on Data Science and Its Applications (ICoDSA)*, July 2024, pp. 346–351.

[22] M. A. Syahira and R. Kurniawan, "Analisis sentimen cyberbullying pada media sosial X menggunakan metode Support Vector Machine," *Jurnal Media Informatika Budidarma*, vol. 8, no. 3, pp. 1724–1733, 2024.

[23] S. Ernawati, F. Frieyadie, and E. R. Yulia, "Cybersentinel: The cyberbullying detection application based on machine learning and vader lexicon with gridsearchcv optimization," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 6, no. 4, pp. 533–542, 2024.

[24] I. S. Arfan, S. Fauziah, and I. Nawangsih, "Analisis sentimen terhadap cyberbullying di X menggunakan algoritma Naive Bayes," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 4, pp. 1411–1419, 2024.

[25] N. Habbat, H. Anoun, and L. Hassouni, "Arabertopic: A neural topic modeling approach for news extraction from arabic facebook pages using pretrained bert transformer model," *International Journal of Computing and Digital System*, vol. 14, 2021.