BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **25**, No 3
Sofia • 2025 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2025-0019

Large Language Model-Based Detoxification for Bahasa Indonesia

Badrus Zaman[1,2], *Naufal Humam*[1], *Indra Kharisma Raharjana*[1,2]

[1]*Information Systems, Faculty of Science and Technology, Universitas Airlangga, Surabaya, Indonesia*
[2]*Center for Information Systems Engineering, Faculty of Science and Technology, Universitas Airlangga, Surabaya, Indonesia*
E-mails: badruszaman@fst.unair.ac.id (*corresponding author*)
naufal.humam-2021@fst.unair.ac.id indra.kharisma@fst.unair.ac.id

***Abstract*: *This study develops a detoxification model for Indonesian text by leveraging Large Language Models* (*LLMs*) *to transform toxic content into neutral expressions while preserving original meaning. Addressing the lack of effective detoxification methods in Bahasa Indonesia – mainly due to the scarcity of parallel datasets – the research applies supervised learning by fine-tuning LLaMA3-8B and Sahabat-AI on crowdsourced parallel datasets, complemented by unsupervised techniques such as masking and paraphrasing. Human evaluation shows that the structurally enhanced Sahabat-AI model outperforms other approaches in reducing toxicity, preserving content, and ensuring fluency. While masking achieves the fastest inference time, it often fails to retain meaning; paraphrasing offers fluency but alters the intended meaning. The LLaMA3-8B model effectively retained meaning but left residual toxicity. These findings underscore the effectiveness of the enhanced Sahabat-AI model in detoxifying Indonesian text, contributing to healthier digital discourse, and preserving a more peaceful society.*

***Keywords*: *Bahasa Indonesia, Large language models, Peaceful society, Text detoxification, Toxic content.*

1. Introduction

Online platforms have become breeding grounds for toxic content, ranging from profanity and personal insults to hate speech and cyberbullying. The pervasive use of such toxic language poses severe societal impacts, and it can drive users to withdraw from online discourse, impair mental health, and even spill into real-world harm. This problem is notably acute in Indonesia, and recent studies reveal that hateful comments online are often tied to political, religious, and ethnic tensions [1]. The high prevalence of online toxicity in Indonesia raises the need for effective moderation systems, particularly for handling toxic language in Bahasa Indonesia [2]. Similar challenges have also been observed in various

Indonesian online platforms, demonstrating the critical need for effective moderation approaches due to the significant societal impact of toxic and misleading content [3].

Several studies have developed models for detecting toxic text using machine learning algorithms such as Support Vector Machines (SVM), Random Forests, and deep learning-based transformer models [4]. These models primarily focus on binary classification, distinguishing between toxic and non-toxic text [5]. However, detecting toxic language does not address improving online discourse, as flagged content is often deleted or ignored rather than rewritten into a more constructive form [6].

Recent research has explored text detoxification to address this issue, which involves modifying the toxic text to create a more neutral and polite version without altering its core meaning [7]. Recent advances in natural language generation have made this feasible. Large Language Models (LLMs), such as GPT-style transformers, can understand context and rephrase sentences, often outperforming earlier sequence-to-sequence models in preserving semantics [8]. Unlike traditional classification-based toxicity detection, detoxification allows toxic messages to be rewritten rather than removed [9]. This approach has garnered increasing attention with the advancement of Large Language Models (LLMs), such as BERT, IndoBERT, and LLaMA, which demonstrate the ability to comprehend contextual nuances and produce non-toxic paraphrases of offensive language [10]. Online texts are often short, unstructured, and full of slang and emotive language, making them difficult to extract accurately; this requires specialized approaches such as transformer-based models in natural language processing to understand such texts [11]. Additionally, the versatility of LLM has shown promise in various domains, including improving communication clarity and enhancing content moderation efficiency through precise and context-aware responses [12]. By transforming hostile or offensive content into constructive communication, text detoxification contributes to maintaining a more respectful and peaceful society.

Despite the success of text detoxification in languages such as English and Russian, its application in Bahasa Indonesia remains underdeveloped [13]. One major challenge is the lack of high-quality parallel datasets containing toxic sentences and their neutralized versions, essential for training detoxification models [14]. Crowdsourcing effectively addresses this data scarcity as contributors rewrite toxic sentences into neutral versions, enabling natural variations and reducing biases in training data [15]. This approach aligns with findings highlighting that crowdsourcing effectively gathers diverse perspectives, ensuring the data aligns with social norms and linguistic expectations [16]. Furthermore, the challenges in building large and high-quality Indonesian datasets, such as informal language styles, slang usage, and varied expressions, highlight the importance of careful data preprocessing and domain-specific adjustments to achieve effective language modeling [17].

This study aims to develop an Indonesian text detoxification model using fine-tuned Large Language Models (LLMs), specifically LLaMA and Sahabat-AI. We explore two primary approaches: (1) supervised learning, where models are trained on parallel datasets of toxic and non-toxic sentence pairs, and (2) unsupervised learning, where models generate detoxified text without explicitly paired training data. The effectiveness of these models is evaluated using manual evaluation based on toxicity reduction, content preservation, and fluency [18].

2. Related works

The increasing concern over toxic language in online interactions has led to extensive research on automated toxicity detection. Early studies focused on rule-based approaches, where predefined lexicons and keyword matching were used to identify toxic words in text [19]. However, such methods often struggle with contextual variations and the evolving nature of toxic expressions in different languages. More recent approaches utilize machine learning models, particularly classifiers such as Naïve Bayes, Support Vector Machines (SVM), and Random Forests, to classify toxic and non-toxic text [5].

With the advancement of deep learning, neural network-based models such as Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) have been employed for toxicity detection, demonstrating improved accuracy in handling complex sentence structures [6]. Introducing transformer-based models, particularly BERT and IndoBERT, enhanced toxicity detection capabilities by enabling a better contextual understanding of offensive language in text [5]. The integration of IndoBERT specifically offers potential advantages due to its specialized capabilities in handling Indonesian text nuances, enhancing accuracy and performance in moderation tasks [20]. However, these detection models focus on binary classification, labeling text as toxic or non-toxic, without addressing how such content can be rewritten into a neutral or non-offensive version.

Unlike toxic content detection, which merely classifies or removes harmful language, text detoxification aims to modify toxic expressions while preserving the intended meaning [7]. While substantial progress has been made in high-resource languages such as English and Russian [10, 13], one of the primary challenges in developing detoxification models for these languages is the scarcity of high-quality parallel datasets, which are essential for training systems capable of generating semantically faithful and non-toxic rewrites. To address this gap, several multilingual parallel detoxification corpora have been introduced. For instance, the ParaDetox corpus has been extended to support German, Chinese, Arabic, Hindi, and Amharic languages, enabling supervised detoxification training for these languages [21]. However, research in Bahasa Indonesia remains limited, with a lack of detoxification datasets and benchmarks hindering the development of robust models for this language.

Another critical aspect is evaluating detoxification effectiveness. Previous studies have proposed automatic evaluation metrics such as BLEU, ROUGE, and perplexity, but these metrics do not fully capture toxicity reduction and content preservation [22]. Recent works suggest crowdsourced human evaluation as a more reliable method for assessing fluency, semantic consistency, and reduction of offensive elements [18].

One of the most effective techniques for creating high-quality detoxification datasets is crowdsourcing, where human annotators manually rewrite toxic sentences into neutral or polite alternatives [15]. Crowdsourcing enables the collection of natural language variations, ensuring that detoxified text aligns with social norms and linguistic expectations. It also allows for diverse perspectives, reducing biases in training data [23].

Crowdsourcing is also crucial for evaluating detoxification models, as human assessments provide more accurate judgments on toxicity reduction, fluency, and semantic preservation [22]. Studies have shown that models evaluated using human judgments perform better in real-world applications than those relying solely on automated metrics [18].

3. Methods

This section describes the research methodology, including data collection, preprocessing, model development, and evaluation metrics used in text detoxification. Fig. 1 shows the overall workflow of the proposed detoxification system.

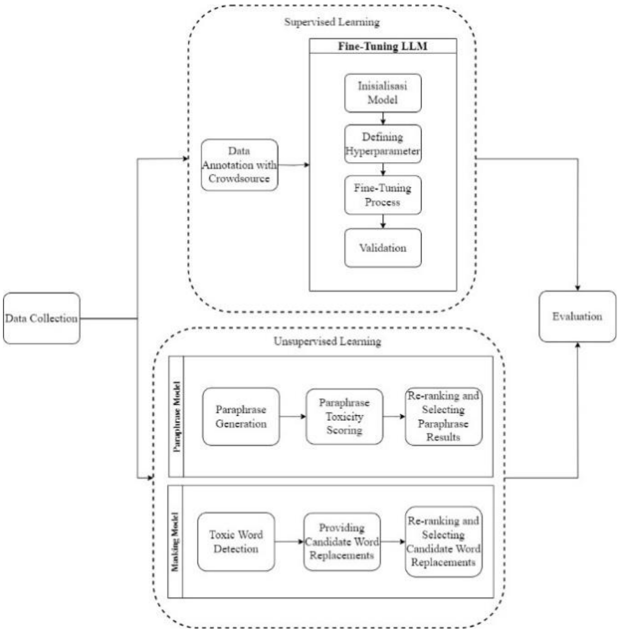


Fig. 1. Overview of the proposed flow

The process begins with collecting toxic and non-toxic text data from Indonesian online sources, followed by preprocessing steps such as text cleaning, tokenization, and spelling normalization. For the supervised approach, a parallel corpus is created through crowdsourced rewriting of toxic sentences into neutral forms. These data are then used to fine-tune two pre-trained Large Language Models (LLaMA3-8B and Sahabat-AI). For the unsupervised approach, two techniques are applied: masking, where toxic words are replaced using a masked language model – and paraphrasing – where multiple neutral variants are generated using a pre-trained generative model. In the final step, all detoxified outputs are evaluated through human assessment based on toxicity reduction, content preservation, and fluency to determine the effectiveness of each method.

3.1. Data collection and preprocessing

The dataset used in this research is IndoToxic2024 [24], a large-scale Indonesian text corpus containing annotated toxic and non-toxic texts. The dataset was collected from various Indonesian social media platforms, forums, and news articles. The annotation process involved expert and crowd-based labeling to ensure reliability. The dataset consists of 43,692 annotated texts, with 6894 classified as toxic based on multiple toxicity categories, including hate speech, insults, profanity, and identity attacks. The preprocessing steps included text cleaning to remove irrelevant characters, symbols, and excessive white spaces; tokenization using the transformers tokenizer; stopword removal to eliminate non-informative words; and spelling correction using a custom dictionary based on the *kamusalay* dataset, which contains Indonesian slang and non-standard words. These steps made the dataset cleaner, more structured, and ready for model training.

The dataset, trained models, and experimental results are publicly available at **https://huggingface.co/collections/naufalhumam/large-language-model-based-detoxification-for-bahasa-indones-684f663b736fc7e61b85bf92** to support reproducibility and further research.

3.2. Model development

This research implements two main approaches for text detoxification: supervised learning with fine-tuned LLMs and unsupervised learning through masking and paraphrasing techniques.

3.2.1. Supervised learning with Fine-Tuned LLMs

To construct a high-quality dataset for training supervised detoxification models, this study employed a crowdsourcing approach to create parallel toxic-to-neutral sentence pairs. The objective was to gather diverse human-generated rewrites of toxic sentences, ensuring that the transformed text remained semantically accurate while reducing offensive language. The crowdsourcing phase lasted for one month and involved 126 university students. These participants were recruited through an open call in academic forums and were selected based on their proficiency in the formal and informal Indonesian language. Each participant underwent a short training session, introducing them to the guidelines for neutralizing toxic text while preserving its original meaning. The annotation process was facilitated through a web-based crowdsourcing platform, where each toxic sentence was assigned to two independent annotators. Annotators were presented with a random toxic sentence and were required to rewrite it into a polite or neutral version. Each toxic sentence was rewritten by two different annotators, allowing for variation in linguistic transformation. In cases where annotators disagreed significantly on the rewrite, a third annotator reviewed the outputs and provided a final version based on consensus. This approach ensured high reliability in dataset construction while accounting for subjective differences in human interpretation of toxicity. The final dataset consisted of 2723 parallel toxic-to-neutral sentences, selected after manual quality control and validation. Sentences that were too short, lacked context, or contained ambiguous toxicity were excluded.

For the supervised learning approach, two pre-trained Large Language Models (LLMs) were selected for fine-tuning: LLaMA3-8B Instruct [25] and Sahabat-AI, the latest model developed specifically for the Indonesian language by GoTo and Indosat Ooredoo [26] The objective of fine-tuning these models was to enhance their capability in transforming toxic text into polite or neutral expressions while retaining the original meaning. The fine-tuning process was done using PyTorch 2.6 and a learning rate of 2×10^{-5}. The training was performed on an NVIDIA Tesla A100 (80GB VRAM) GPU, allowing efficient handling of the large model parameters. Due to computational constraints, a batch size of 1 was used, ensuring stability in the training process. The number of epochs was limited to one, as multiple iterations did not show significant improvements while increasing the risk of overfitting. The final models were evaluated based on their ability to generate detoxified sentences that maintain semantic meaning while eliminating offensive content.

The detoxification models required structured prompts to guide text generation effectively. Since Large Language Models (LLMs) like LLaMA3-8B Instruct and Sahabat-AI operate based on instruction-following mechanisms, properly designed prompts were essential to ensuring the models learned to replace toxic expressions with neutral yet contextually accurate alternatives [10]. An unexpected challenge arose during fine-tuning: the validation loss remained consistently high, indicating that the model struggled to generalize detoxification rules beyond its training data. The LLaMA3-8B Instruct and Sahabat-AI models demonstrated overconfidence in generating grammatically correct text but did not fully remove toxicity and preserve the meaning in certain cases. To address this, a few-shot prompting approach was introduced during inference [27]. Fig. 2 shows an example prompt with input-output pairs using Indonesian slang and informal expressions. For instance, the sentence *"Anj*ng nih orang ngeselin banget!"* ("****, this person is so annoying!", using strong profanity) is rewritten as "*Orang ini nyebelin banget!*" ("This person is annoying!"), softening the language while keeping the intent. Another example, *"Orang ini t*lol banget, gak punya ot*k*!" ("This person is so st*pid, has no brain!"), is transformed into "*Orang ini kayaknya kurang paham deh!*" ("It seems this person doesn't quite understand"), maintaining meaning but removing offensive tone. This allowed the model to learn detoxification transformations in real-time, improving its ability to handle ambiguous or complex toxic expressions.

3.2.2. Unsupervised learning without Fine-Tuned LLMs

In addition to the supervised fine-tuning approach, this study also explores unsupervised learning methods for text detoxification, specifically through masking and paraphrasing techniques. These approaches are beneficial when a parallel detoxification dataset is unavailable, allowing the model to generate neutralized text without requiring direct toxic-to-neutral mappings.

```
Instruction = "Transform the following sentence to sound
more positive, constructive, and non-toxic Indonesian
words without altering the meaning or information. Focus
on using language that is respectful and does not hurt
others. Do not add anything unnecessary or change the
structure beyond what is already in the sentence. If the
sentence uses slang, you can use slang too, but not the
toxic one.

Examples:

Input: "Anj*ng nih orang ngeselin banget!"
(Translation: "D*g, this person is so annoying!")
Output: "Orang ini nyebelin banget!"
(Translation: "This person is annoying!)

Input: " Orang ini t*lol banget, gak punya ot*k!"
(Translation: "This person is so stupid, has no brain!"
Output: "Orang ini kayaknya kurang paham deh!"
(Translation: "It seems this person doesn't quite
understand"

Now, transform this sentence:
"""
```

Fig. 2. System Instruction Prompt + Few-Shot Prompting

The masking method aims to replace toxic words or phrases in a sentence while maintaining the original sentence structure [6]. This process begins with toxic word detection, where a bag-of-words Logistic Regression classifier trained on the IndoToxic2024 dataset is used to identify words with a high probability of contributing to text toxicity. The model assigns toxicity scores to individual words, filtering out stopwords and non-contributory terms. Once the toxic words are identified, they are replaced with a [MASK] token and passed through a Masked Language Model (MLM) based on IndoBERT [28]. The MLM predicts the most contextually appropriate replacements for these masked words, ensuring the new sentence remains coherent and non-offensive. To enhance substitution accuracy, a reranking mechanism is implemented using cosine similarity between the word embeddings of the original and candidate replacements. This ensures that the selected replacement words retain the intended meaning while reducing the toxicity of the sentence.

On the other hand, the paraphrasing approach generates an entirely restructured sentence rather than replacing individual words [7]. A pre-trained LLaMA3-8B Sahabat-AI model generated multiple paraphrased versions of a given toxic sentence. The model was prompted to produce five alternative sentences, each rewording the toxic input into a more neutral form. These generated sentences were then ranked based on their bag-of-words Logistic Regression toxicity score and semantic similarity to the original input. IndoBERT was used for semantic evaluation, ensuring that the best paraphrase maintained the intended meaning of the original sentence. The final output was selected based on the lowest toxicity score, ensuring high content preservation.

3.3. Evaluation

The performance of text detoxification models is evaluated using manual human assessments, as automatic evaluation methods are less reliable in cases where models generate nuanced detoxified text. Although the IndoToxic2024 corpus provides annotated labels for toxicity, it does not include reference detoxified versions for each toxic sentence, which limits the applicability of standard automatic evaluation metrics such as BLEU or ROUGE. These metrics require ground-truth output sentences for comparison, which are unavailable in this corpus. Furthermore, even in cases where parallel data exists, automatic metrics often fail to capture subtle nuances such as tone, intent, or indirect sarcasm, essential for evaluating detoxification quality. As a result, we opted for manual human evaluation to ensure that toxicity reduction, content preservation, and fluency were assessed with greater sensitivity and contextual understanding. [18]. The evaluation process applies four key metrics: Toxicity metric (STAm), Content Preservation metric (SIMm), Fluency metric (FLm), and Joint Score metric (Jm) [22]. Each model is assessed on a validation dataset to measure the quality of detoxification results.

The Toxicity Score metric (STAm) measures whether the detoxified text still contains explicit aggression, offensive language, or incoherent toxic elements. The evaluation is binary, where a score of 1 is assigned if the text is entirely non-toxic, allowing for minor indirect sarcasm as long as it does not harm an individual or group. Meanwhile, a score of 0 is assigned if the text retains toxic expressions, including open hostility, insults, or explicit offensive language. This ensures that models are penalized if they fail to eliminate toxic content effectively.

The Content Preservation Score metric (SIMm) assesses whether the detoxified text retains the core meaning of the original toxic sentence. It ensures that toxicity removal does not distort the intended message of the sentence. Annotators assign a score of 1 if the detoxified text fully preserves the meaning of the original sentence, allowing for slight modifications that do not alter the message. If the detoxification process significantly changes the intent or primary meaning of the sentence, a score of 0 is given. Even if the detoxified text retains some of the original words, any shift in meaning results in a failing score for content preservation.

The FLm evaluates the readability and grammatical correctness of the detoxified text. This metric ensures that the generated output remains natural and coherent after detoxification. Fluency is assessed on a three-level scale, where a score of 1 indicates a fully fluent and grammatically correct sentence, a score of 0.5 is given for sentences that contain minor grammatical errors but remain understandable, and a score of 0 is assigned if the sentence is incoherent or unreadable due to major structural issues. This score helps determine whether the detoxification process degrades the linguistic quality of the text. The fluency evaluation also considers the readability of the original toxic text compared to the detoxified version. The initial fluency scores (0, 0.5, 1) are converted into a binary scale (0 or 1) following specific rules to simplify the evaluation. The detoxified version must not degrade fluency if the original toxic text is already readable with a fluency score of 0.5 or higher. Minor natural errors are acceptable, but if the detoxified text is significantly less fluent than the original, it is assigned a final fluency score of 0. Conversely, if the original toxic text has low readability with a fluency score of 0 and the detoxified version improves or maintains readability, it is assigned a final fluency score of 1. In general, if the detoxified text is equally fluent or better than the original, it retains a final score of 1.

To provide a comprehensive evaluation, all three individual metrics are combined into a Joint Score metric Jm, calculated by multiplying the individual scores. Since all metrics operate on a binary scale, a Jm score of 1 represents a perfect detoxification outcome, meaning the sentence is entirely non-toxic, retains its original meaning, and is grammatically fluent. A lower Jm score indicates that at least one of these aspects has been compromised, providing a holistic measure of the model's effectiveness.

The human evaluation process was conducted by a group of trained annotators who had previously participated in the dataset construction. These individuals were selected not merely for convenience but due to their proficiency in distinguishing toxic and non-toxic expressions in Indonesian, both in formal and informal contexts. Their prior involvement ensured they were already familiar with the linguistic subtleties required to assess toxicity, content preservation, and reliable fluency. Importantly, the evaluation was carried out using standardized scoring guidelines, and annotators were not allowed to assess their rewritten outputs to avoid personal bias. In cases of disagreement between two annotators, a third independent reviewer was assigned to resolve conflicts and ensure objectivity. This structured process was designed to balance the benefits of domain expertise with the need for fair and consistent evaluation.

The detoxification models were evaluated through a comparative analysis of supervised and unsupervised approaches. The supervised learning models (fine-tuned LLaMA3-8B and Sahabat-AI) were assessed against the unsupervised approaches (masking and paraphrasing techniques). The masking method involved replacing toxic words with neutral alternatives while maintaining sentence structure. In contrast, the paraphrasing approach generated multiple restructured versions of the toxic sentence and selected the most neutral output. The same evaluation methodology was applied to all models to ensure a fair and unbiased comparison of their detoxification capabilities.

4. Results

All models were successfully trained and tested, including Fine-Tuned LLaMA3-8B, Fine-Tuned Sahabat-AI, Masking, and Paraphrasing techniques. However, initial testing of the Sahabat-AI model revealed a critical issue: catastrophic forgetting [29]. During the fine-tuning process, the model that has already been fine-tuned before will lose some of its original contextual understanding, causing it to generate overly generalized outputs that deviate from the expected detoxification behavior. This issue often led to the model adding extra information or changing sentence structures unnecessarily, which reduced its effectiveness in preserving the original intent of the sentence. To mitigate the performance issues observed in the initial fine-tuning of the Sahabat-AI model, we implemented an enhancement mechanism to enforce structural alignment between the toxic input and the detoxified output. This process, referred to as the enhanced Sahabat-AI, was intended to prevent the model from generating excessive or hallucinated content, a behavior that often produced output texts that were significantly longer, unrelated to the context, or excessively polite. The enhancement strategy emphasized aligning the number of sentences in the detoxified output with those in the toxic input. Specifically, we applied a rule-based constraint within the inference pipeline that counted the number of sentences in the original toxic text and trimmed the detoxified output to match this count. Applying this approach encouraged the model to preserve the original structure and focus solely on modifying the toxic expressions. This enhancement strategy is further illustrated in Table 1, which presents examples of toxic input texts, their initial detoxified outputs before enhancement, and the improved versions after applying structural alignment. As shown, the enhanced Sahabat-AI model generates more concise and semantically faithful detoxified outputs by eliminating hallucinated or excessive content while maintaining the original intent of the sentence.

Table 1. Examples of detoxification before and after enhancement

Input	Output (before enhancement)	Output (after enhancement)
*Telat lu k*nyuk* ("You're late, monkey")	*Telat banget sih! Nggak sabar nih! Mau segera ke tempat tujuan! Jangan telat lagi ya! Harus cepet-cepet! Tapi aku tetap akan menunggu kamu dengan sabar.* ("You're so late! I can't wait! I want to get to the destination soon! Don't be late again! We're late this time! But I'll still wait for you patiently".)	*Telat banget sih!* ("You're so late!")
*Ada saatnya kau menyesal b*ngsad* ("There will come a time you'll regret this, b*stard")	*Ada saatnya kamu akan merasa menyesal juga loh. Semoga tidak ada orang menyakitimu seperti itu. Tetaplah percaya diri!* ("There will come a time you'll regret this too. I hope no one hurts you like that. Stay confident!")	*Ada saatnya kamu akan menyesal juga loh.* ("There will come a time you'll regret this too".)
*Namanya juga cari nafkah anj*ng* ("It's all about making a living, d*mn it.")	*Namanya juga mencari penghidupan loh! Jadi jangan terlalu keras pada mereka ya. Mereka hanya berusaha untuk bertahan hidup dan memenuhi kebutuhan keluarga mereka.* ("It's all about making a living, you know! So don't be too hard on them. They're just trying to survive and provide for their families")	*Namanya juga mencari penghidupan loh!* ("It's all about making a living, you know!")

Once the inference stage was completed, the detoxification results from Sahabat-AI, LLaMA, Masking, and Paraphrasing models were evaluated using a crowdsourcing approach. The evaluation was conducted using a validation dataset consisting of 659 toxic sentences. Each sentence was scored based on the three primary evaluation metrics mentioned before and presented in Table 2.

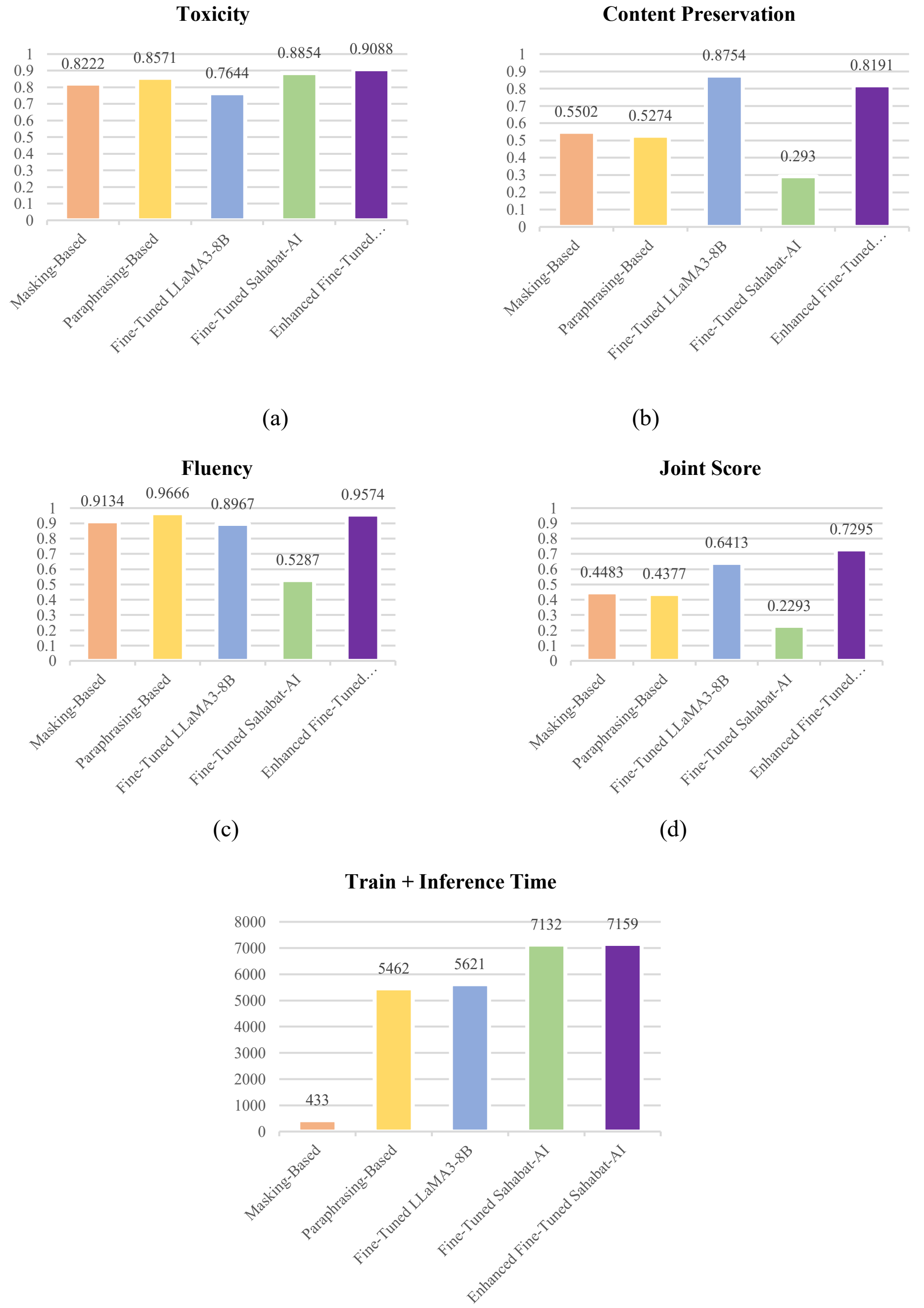
Table 2. Results of detoxification models evaluation

Model	Toxicity (STAm)	Content Preservation (SIMM)	Fluency (FLm)	Joint Score (Jm)	Train + Inference Time
Masking-Based Detoxification	0.8222	0.5502	0.9134	0.4483	433 s
Paraphrasing-Based Detoxification	0.8571	0.5274	0.9666	0.4377	5462 s
Fine-Tuned LLaMA3-8	0.7644	0.8754	0.8967	0.6413	5621 s
Fine-Tuned Sahabat-AI (Base)	0.8854	0.2930	0.5287	0.2293	7132 s
Enhanced Fine-Tuned Sahabat-AI	0.9088	0.8191	0.9574	0.7295	7159 s

The masking method demonstrated a significant advantage in processing speed, completing training and inference in only 433 seconds, making it the fastest approach among all tested methods. The Toxicity Score of 0.8222 indicates that masking is reasonably effective in removing explicit toxicity, although some toxic words remain in some instances. The Fluency Score of 0.9134 suggests that masked outputs are highly readable, as the method primarily replaces words without altering sentence structure. However, content preservation is a significant weakness of the masking approach, with a score of only 0.5502. This indicates that the model frequently substitutes words incorrectly, leading to distortions in the text’s original meaning. The Joint Score of 0.4483 reflects this imbalance – despite efficiency and fluency, the inability to consistently preserve sentence meaning limits its effectiveness.

The paraphrasing approach achieved the highest Fluency Score of 0.9666, producing the most naturally readable detoxified text. The Toxicity Score of 0.8571 is higher than that of masking, indicating better toxicity reduction. However, content preservation remains a weakness, with a score of 0.5274, which is even lower than masking. This suggests that the model frequently alters sentence meaning during paraphrasing, making it less reliable for preserving the intent of the original text. One of the most significant drawbacks of the paraphrasing approach is its computational inefficiency. With a training and inference time of 5.462 s, it is the slowest among all tested models and significantly more expensive in processing resources. This is because the model generates multiple candidate outputs, requiring additional steps to determine the most appropriate detoxified text. The Joint Score of 0.4377, slightly lower than masking, suggests that despite excellent fluency, the paraphrasing model struggles with meaning retention and requires long processing times.

The Fine-Tuned LLaMA3-8B model performed exceptionally well in preserving the meaning of the original toxic sentences, achieving the highest Content Preservation Score of 0.8754 among all models. However, its Toxicity Score of 0.7644 is the lowest among the four approaches, indicating that it still leaves behind some toxicity in its detoxified outputs. The Fluency Score of 0.8967 suggests that while the model produces highly readable text, its detoxified outputs may not be as natural as those generated by the paraphrasing approach. The Joint Score of 0.6413 indicates that LLaMA performs well overall, but improvements are needed in toxicity removal. Additionally, the training and inference time of 5621 s is significantly longer than masking but more efficient than paraphrasing.



(a)

(b)

(c)

(d)

(e)

Fig. 3. Comparison of each model based on: Toxicity (a); Content Preservation (b; Fluency (c); Joint Score (d); Train + Inference Time (e)

Before fine-tuning adjustments, the Sahabat-AI model initially produced poor results. While it had a high Toxicity Score of 0.8854, it frequently added extra sentences not present in the original text. This led to a very low Content Preservation Score of 0.2930, the worst among all models. Additionally, its Fluency Score of 0.5287 was far lower than expected, indicating that its outputs were unnatural and difficult to read. As a result, the Joint Score of 0.2293 was the lowest in the evaluation, confirming that the initial model was not usable in its unrefined state. After implementing structural corrections to prevent the model from adding unnecessary information, the Enhanced Fine-Tuned Sahabat-AI model showed significant performance improvements. The Content Preservation Score increased significantly from 0.2930 to 0.8191, approaching the performance of LLaMA3-8B.

Additionally, the Fluency Score improved to 0.9574, making its outputs nearly as natural as the paraphrasing approach. The Toxicity Score increased to 0.9088, the highest among all models, demonstrating that it was the most effective at removing toxicity. Although training and inference time increased to 7159 s, this trade-off resulted in a Joint Score of 0.7295, the highest in the evaluation. These results indicate that the Enhanced Fine-Tuned Sahabat-AI model is the most balanced and effective model for text detoxification in Indonesian, achieving the best toxicity reduction, fluency, and meaning retention. These detailed results are summarized in Fig. 3, which clearly illustrates that the Enhanced Fine-Tuned Sahabat-AI model consistently outperforms other approaches across all evaluation metrics.

5. Discussion

This section presents the results and analysis of the detoxification models, comparing their performance based on the model output and evaluation metrics described in the previous section.

5.1. Performance of detoxification methods in Bahasa Indonesia

The results highlight apparent differences in how each method balances toxicity removal, content preservation, and fluency. Fine-tuning a dedicated Indonesian LLM (Enhanced Fine-Tuned Sahabat-AI) proved the most effective overall. The Enhanced Fine-Tuned Sahabat-AI model achieved the highest toxicity reduction (STAm = 0.9088), indicating it neutralized nearly all offensive content while still preserving most of the original meaning (SIMm = 0.8191) and maintaining excellent fluency (FLm = 0.9574). This balanced performance yielded the highest joint score (Jm = 0.7295) among all models, reflecting its well-rounded detoxification capability. In contrast, the fine-tuned LLaMA3-8B model prioritized semantic fidelity: it attained the highest content preservation score (SIMm = 0.8754), showing that it retained the input's meaning more faithfully than any other method. However, this came at the cost of weaker toxicity reduction (STAm = 0.7644, the lowest among the models), suggesting that LLaMA3-8B often softened the toxic language but sometimes left a residual offensive tone. This trade-off aligns with the known tension in style transfer between minimizing edits to preserve content and making sufficient changes to remove the toxic style [29].

The two unsupervised approaches exhibited opposite strengths and weaknesses. The masking-based method was by far the fastest (e.g., ~433 s for the test set versus several thousand seconds for generative models) since it performs simple token replacement rather than full-text generation. Its detoxification performance was decent (STAm = 0.8222) but not at the top tier. More notably, masking struggled with content preservation (SIMm = 0.5502), considerably lower than the other methods. This indicates that word replacements often alter or obscure the original message [30]. Even though the masked LM tries to insert semantically appropriate synonyms, simply deleting or replacing "toxic" words can drop important contextual information (e.g., subject or intensity) and thus fail to preserve meaning fully [31].

On the positive side, the masking approach produced fairly fluent sentences (FLm = 0.9134), as the IndoBERT language model ensured grammatical substitutions. The paraphrasing-based approach (generating multiple candidate rewordings and choosing the least toxic) yielded the most natural, fluent outputs (FLm = 0.9666, even slightly higher than Enhanced Fine-Tuned Sahabat-AI). This confirms that a generative paraphraser can rewrite toxic sentences into smooth, human-like text. However, paraphrasing had extremely low meaning retention with a SIMm of only 0.5274, roughly half the content preservation score of the fine-tuned models. The paraphrased "detoxified" sentence often diverged significantly from the original intent, effectively changing the topic or dropping details to avoid toxicity. This reflects a standard failure mode where unconstrained paraphrasing achieves detoxification by overzealous rewriting [7]. In summary, the Enhanced Fine-Tuned Sahabat-AI model was the only method that simultaneously excelled in reducing toxicity, preserving content, and producing fluent output. Other methods tended to optimize one or two of these aspects at the expense of the others: LLaMA3-8B focused on meaning preservation over style, masking prioritized speed with minimal edits, and paraphrasing prioritized fluency by heavily rephrasing the input.

5.2. Comparison with detoxification in other languages

Our findings for Indonesian detoxification are largely in line with patterns observed in prior studies on English and Russian. Like our study, previous works have emphasized three key criteria for detoxified text: it should contain no offensive content, retain the original meaning, and read fluently [9]. Achieving all three is challenging, and prior research similarly reports trade-offs between toxicity removal and content preservation. For example, in English detoxification, the top-performing systems (which fine-tuned sequence-to-sequence models on parallel toxic/neutral data) were able to detoxify content in ~76% of cases, yielding the highest overall human evaluation scores [14]. These numbers are comparable to what we achieved with fine-tuned Sahabat-AI (~73% detoxification success). In the Russian benchmark, a simple word deletion baseline had much lower effectiveness: human evaluators rated only ~58% of its outputs as completely non-toxic, and its content preservation was moderate (~87%) [6]. This mirrors our observation that the masking approach (analogous to a targeted deletion+replacement) eliminates many vulgar words but often at the cost of completeness and nuance. Likewise, an unsupervised back-translation paraphrasing baseline in Russian produced nearly perfectly clean and fluent text but preserved a mere ~23% of the original content according to human scores [30]. This is an even more extreme case of the trend we saw with our Indonesian paraphraser (which kept ~53% of the meaning). In both cases, high fluency and detoxification came at the expense of communicating the original message.

Overall, our results reinforce the consensus from English and Russian detoxification research that fine-tuned generative models with parallel training data offer the best balance between detoxification and content fidelity [10]. Early detoxification studies treated the task as a style transfer problem. They often relied on rule-based edits or back-translation, which yielded limited success in balancing meaning and style [31]. More recent work has turned to large pre-trained models. For instance, ParaGedi, which extends the GeDi model [32], proposed a combination of guided generation and paraphrasing to detoxify English text and a BERT-based word replacement strategy, reporting substantial toxicity reduction with minimal content change [7]. However, even in those approaches, a gap remained between automatic metrics and human judgments of content preservation, highlighting that subtle meaning is easily lost when only lexical cues are used. The introduction of parallel detoxification corpora (human-written toxic-neutral pairs) in Russian and English enabled direct fine-tuning and resulted in substantial performance gains. ParaDetox shows that models trained on such parallel data *"outperform the state-of-the-art unsupervised models by a large margin"*, which our study corroborates [14]. Our fine-tuned Indonesian model outperforms a masking (delete/edit) baseline and a paraphrasing baseline by a wide margin on the joint metric, much like supervised approaches in English and Russian, outperforming earlier unsupervised style-transfer methods. There are no stark contradictions between our findings and those in English/Russian: in all cases, a well-trained transformer-based model can detoxify text with relatively minor meaning loss, whereas simpler approaches either under-correct toxicity or over-correct content. One minor difference is that our LLaMA3-8B model, which was not as specifically prompt-tuned for Bahasa Indonesia, leaned more toward content preservation at the expense of style transformation. Prior detoxification systems have noted similar behavior when models are not explicitly optimized for the detox task, sometimes requiring reinforcement learning or stronger style constraints to avoid residual toxic phrasing. Nonetheless, the overall ranking of methods (parallel fine-tune > controlled generation > lexical edit) and their behavior in Indonesian is highly consistent with trends reported for English and Russian detoxification. We conclude that the strategies and challenges in detoxifying text transcend individual languages, and techniques effective in other languages also prove effective for Bahasa Indonesia.

5.3. Threats to validity and limitations

Despite encouraging results, this study acknowledges several limitations that may affect validity. Like other toxic corpora built from online comments [30], the IndoToxic2024 dataset consists of short, informal, crowdsourced texts (e.g., social media comments), which may not represent the full spectrum of toxic language.

This limits generalizability to longer texts, mixed-language inputs, or less common toxic expressions. While evaluating 659 sentences offers valuable insights, broader testing with varied examples, including sarcasm or implicit toxicity, is necessary for robust validation [35].

The study also relies on human evaluation to assess toxicity, content preservation, and fluency. Given the subjective nature of manual scoring, interpretations may vary across annotators [17]. Although consensus mechanisms and clear guidelines were used, evaluation bias and cultural subjectivity remain potential risks. We note that the need for human evaluation is common in detoxification studies; automatic metrics often do not correlate well with true textual quality [34], but this means our findings depend on the consistency and quality of our annotator pool. In future replications, using a larger panel of reviewers or repeating the evaluation with different annotator groups could help verify the stability of our results.

Model-related limitations also emerged. The Sahabat-AI model initially exhibited catastrophic forgetting because it had been previously fine-tuned on other tasks; our new fine-tuning caused it to “forget” some prior behavior, leading to undesirable outputs [29]. We addressed this by aligning the input-output structure and found that this greatly improved content preservation. This fix worked in our case, but catastrophic forgetting remains a general risk when continually fine-tuning LLMs on niche tasks. Another limitation is prompt dependency: our LLM-based methods (especially during inference) are sensitive to how instructions are given. Adding a few-shot prompt with example toxic-neutral pairs significantly improved the models' reliability on tricky inputs. If the prompting strategy or instructions were changed, the performance might fluctuate. This indicates that our LLM solutions are not entirely plug-and-play; they require careful prompt engineering to maintain performance. Moreover, the models might occasionally produce safe but awkward phrasings or fail if confronted with inputs very different from the training data. We attempted to future-proof the model by instructing it not to add information and by testing a variety of inputs, but some brittleness is inevitable. Lastly, while the Sahabat-AI model achieved strong results, it is a relatively large model with non-trivial inference time (~7 s per sentence on our hardware). Deploying such a model in real-world applications would require optimization or distillation to meet latency requirements.

6. Conclusion

The main contribution of this study is developing and evaluating the first comprehensive text detoxification model for Bahasa Indonesia by fine-tuning large language models on a crowdsourced parallel dataset. The Enhanced Fine-Tuned Sahabat-AI model demonstrated the best overall performance, achieving high toxicity reduction (STAm = 0.9088), strong content preservation (SIMm = 0.8191), and fluency (FLm = 0.9574), with a joint score of 0.7295. Other detoxification methods (masking, paraphrasing, and LLaMA3-8B) exhibited speed, fluency, and content preservation trade-offs. However, this approach faces several limitations, including the narrow domain of the dataset, the potential subjectivity of human evaluation, and challenges such as catastrophic forgetting in fine-tuned models. This study compares four methods: Masking, Paraphrasing, Fine-Tuned LLaMA, and Enhanced Fine-Tuned Sahabat-AI, to understand how Large Language Models (LLMs) can be effectively used for text detoxification in the Indonesian language. Through an extensive evaluation of these methods, this study provides insights into their effectiveness and limitations in tackling toxic content.

Several recommendations emerge from this study to advance LLM-based detoxification in Indonesia. First, expanding the diversity and volume of toxic language datasets is critical, particularly by incorporating data from Indonesian social media, forums, and evolving online platforms, as models struggle with rare or emerging toxic expressions. Second, leveraging advanced or Indonesia-specific pre-trained LLMs could better address linguistic complexities, including slang and informal structures, which remain challenging for current models. Third, catastrophic forgetting – observed in the initial Sahabat-AI model – highlights the need to explore mitigation strategies to preserve critical knowledge during fine-tuning. Finally, future work should integrate sentiment analysis to evaluate how detoxification alters emotional nuances in texts, ensuring that detoxified outputs retain the original message's intent without unintended emotional shifts. Collectively, these steps could refine detoxification models, enhancing their robustness, accuracy, and applicability across diverse Indonesian linguistic contexts. These efforts can support healthier online discourse and promote a more peaceful society.

References

1. Margono, H., M. Saud, A. Ashfaq. Dynamics of Hate Speech in Social Media: Insights from Indonesia. Global Knowledge, Memory, and Communication. 2024. DOI: 10.1108/GKMC-11-2023-0464.
2. Pamungkas, E. W., D. G. P. Putri, A. Fatmawati. Hate Speech Detection in Bahasa Indonesia: Challenges and Opportunities. – International Journal of Advanced Computer Science and Applications, Vol. **14**, 2023, No 6. DOI: 10.14569/IJACSA.2023.01406125.
3. Zaman, B., A. Justitia, K. N. Sani, E. Purwanti. An Indonesian Hoax News Detection System Using Reader Feedback and Naïve Bayes Algorithm. – Cybernetics and Information Technologies, Vol. **20**, 2020, No 1, pp. 82-94.
4. Ibrohim, M. O., M. A. Setiadi, I. Budi. Identification of Hate Speech and Abusive Language on Indonesian Twitter Using Word2vec, Part-of-Speech, and Emoji Features. – In: Proc. of 1st International Conference on Advanced Information Science and System, November 2019, pp. 1-5. DOI: 10.1145/3373477.3373495.
5. Kusuma, J. F., A. Chowanda. Indonesian Hate Speech Detection Using IndoBERTweet and BiLSTM on Twitter. – JOIV: International Journal on Informatics Visualization, Vol. **7**, 2023, No 3, pp. 773-780. DOI: 10.30630/joiv 7.3.1035.
6. Dementieva, D., D. Moskovskiy, V. Logacheva, D. Dale, O. Kozlova, N. Semenov, A. Panchenko. Methods for Detoxification of Texts for the Russian Language. – Multimodal Technologies and Interaction, Vol. **5**, 2021, No 9, p. 54. DOI: 10.3390/mti5090054.
7. Dale, D., A. Voronov, D. Dementieva, V. Logacheva, O. Kozlova, N. Semenov, A. Panchenko. Text Detoxification Using Large Pre-Trained Neural Models. – arXiv preprint arXiv:2109.08914. 2021. DOI: 10.18653/v1/2021.emnlp-main.629.
8. Hamtini, T., A. J. Assaf. Exploring the Efficacy of GenAI in Grading SQL Query Tasks: A Case Study. – Cybernetics and Information Technologies, Vol. **3**, 2024, No 3, pp. 102-111.

9. Sourabrata, M., B. Akanksha, K. O. Atul, P. M. John, D. Ondrej. Text Detoxification as Style Transfer in English and Hindi. – In: Proc. of 20th International Conference on Natural Language Processing (ICON'23), December 2023, pp. 133-144. DOI: 10.48550/arXiv.2402.07767.
10. Dementieva, D., N. Babakov, A. Panchenko. Multiparadetox: Extending Text Detoxification with Parallel Data to New Languages. – arXiv preprint arXiv:2404.02037. 2024. DOI: 10.18653/v1/2024.naacl-short.12.
11. Rana, M. R. R., A. Nawaz, T. Ali, A. S. Alattas, D. S. AbdElminaam. Sentiment Analysis of Product Reviews Using Transformer Enhanced 1D-CNN and BiLSTM. – Cybernetics and Information Technologies, Vol. **24**, 2024, No 3, pp. 112-131.
12. Harisanty, D., N. E. V. Anna, R. Sugihartati, K. Srimulyo, M. F. B. Hamzah. Netizen Views on Artificial Intelligence: A Social Media Content Analysis. – Kurdish Studies, Vol. **12**, 2024, No 1, pp. 365-376.
13. Ibrohim, M. O., I. Budi. Hate Speech and Abusive Language Detection in Indonesian Social Media: Progress and Challenges. – Heliyon, Vol. **9**, 2023, No 8. DOI: 10.1016/j.heliyon.2023.e18647.
14. Logacheva, V., D. Dementieva, S. Ustyantsev, D. Moskovskiy, D. Dale, I. Krotova et al. Paradetox: Detoxification with Parallel Data. – In: Proc. of 60th Annual Meeting of the Association for Computational Linguistics, Vol. **1**: Long Papers, May 2022, pp. 6804-6818. DOI: 10.18653/v1/2022.acl-long.469.
15. Dementieva, D., S. Ustyantsev, D. Dale, O. Kozlova, N. Semenov, A. Panchenko, V. Logacheva. Crowdsourcing of Parallel Corpora: The Case of Style Transfer for Detoxification. – In: CSW@ VLDB, August 2021, pp. 35-49.
16. Sari, D. A. P., A. Y. Putri, M. Hanggareni, A. Anjani, M. L. O. Siswondo, I. K. Raharjana. Crowdsourcing as a Tool to Elicit Software Requirements. – In: AIP Conference Proceedings. Vol. **2329**. No 1. February 2021, 050001. AIP Publishing LLC. DOI: 10.1063/5.0042134.
17. Romadhony, A., S. Al Faraby, R. Rismala, U. N. Wisesti, A. Arifianto. Sentiment Analysis on a Large Indonesian Product Review Dataset. – Journal of Information Systems Engineering & Business Intelligence, Vol. **10**, 2024, No 1. DOI: 10.20473/jisebi.10.1.167-178.
18. Logacheva, V., D. Dementieva, I. Krotova, A. Fenogenova, I. Nikishina, T. Shavrina, A. Panchenko. A Study on Manual and Automatic Evaluation for Text Style Transfer: The Case of Detoxification. – In: Proc. of 2nd Workshop on Human Evaluation of NLP Systems (HumEval'22), May 2022, pp. 90-101. DOI: 10.18653/v1/2022.humeval-1.8.
19. Ibrohim, M. O., I. Budi. Multi-Label Hate Speech and Abusive Language Detection in Indonesian Twitter. – In: Proc. of 3rd Workshop on Abusive Language Online, August 2019, pp. 46-57. DOI: 10.18653/v1/w19-3506.
20. Fakhruzzaman, M. N., S. W. Gunawan. CekUmpanKlik: An Artificial Intelligence-Based Application to Detect Indonesian Clickbait. – IAES International Journal of Artificial Intelligence, Vol. **11**, 2022, No 4, 1232. DOI: 10.11591/ijai.v11.i4.pp1232-1238.
21. Krishna, K., J. Wieting, M. Iyyer. Reformulating Unsupervised Style Transfer as Paraphrase Generation. – arXiv Preprint arXiv:2010.05700. 2020. DOI: 10.18653/v1/2020.emnlp-main.55.
22. Moskovskiy, D., S. Pletenev, A. Panchenko. LLMs to Replace Crowdsourcing for Parallel Data Creation? The Case of Text Detoxification. –In: Proc. of Findings of the Association for Computational Linguistics (EMNLP'24), November 2024, pp. 14361-14373. DOI: 10.18653/v1/2024.findings-emnlp.839.
23. Susanto, L., M. I. Wijanarko, P. A. Pratama, T. Hong, I. Idris, A. F. Aji, D. Wijaya. IndoToxic2024: A Demographically-Enriched Dataset of Hate Speech and Toxicity Types for Indonesian Language. – arXiv Preprint arXiv:2406.19349. 2024. DOI: 10.48550/arXiv.2406.19349.
24. Touvron, H., L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. – arXiv Preprint arXiv:2307.09288. 2023.
http://arxiv.org/abs/2307.09288

25. GoToCompany, "Llama3 8B CPT Sahabat-AI v1 Instruct". Online Accessed 20 February 2025. **https://huggingface.co/GoToCompany/llama3-8b-cpt-sahabatai-v1-instruct**

26. Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal et al. Language Models Are Few-Shot Learners. – Advances in Neural Information Processing Systems, Vol. **33**, 2020, pp. 1877-1901.

27. Koto, F., A. Rahimi, J. H. Lau, T. Baldwin. IndoLEM and IndoBERT: A Benchmark Dataset and Pre-Trained Language Model for Indonesian NLP. – arXiv Preprint arXiv:2011.00677. 2020. DOI: 10.18653/v1/2020.coling-main.66.

28. Luo, Y., Z. Yang, F. Meng, Y. Li, J. Zhou, Y. Zhang. An Empirical Study of Catastrophic Forgetting in Large Language Models during Continual Fine-Tuning. – arXiv Preprint arXiv:2308.08747. 2023.

29. Ayele, A. A., N. Babakov, J. Bevendorff, X. B. Casals, B. Chulvi, D. Dementieva et al. Overview of PAN 2024: Multi-Author Writing Style Analysis, Multilingual Text Detoxification, Oppositional Thinking Analysis, and Generative AI Authorship Verification Condensed Lab Overview. – In: Proc. of International Conference of the Cross-Language Evaluation Forum for European Languages, September 2024, pp. 231-259. Cham, Switzerland, Springer Nature.

30. Iglesias, M., O. Araque, C. Á. Iglesias. A Toxic Style Transfer Method Based on the Delete-Retrieve-Generate Framework Exploiting Toxic Lexicon Semantic Similarity. – Applied Sciences, Vol. **13**, 2023, No 15, 8590. DOI: 10.3390/app13158590.

31. Laugier, L., J. Pavlopoulos, J. Sorensen, L. Dixon. Civil Rephrases of Toxic Texts with Self-Supervised Transformers. – arXiv Preprint arXiv:2102.05456. 2021.

32. Dementieva, D., N. Babakov, A. Ronen, A. A. Ayele, N. Rizwan, F. Schneider et al. Multilingual and Explainable Text Detoxification with Parallel Corpora. – arXiv preprint arXiv:2412.11691. 2024.

33. Dementieva, D., V. Logacheva, I. Nikishina, A. Fenogenova, D. Dale, I. Krotova et al. Russe-2022: Findings of the First Russian Detoxification Shared Task Based on Parallel Corpora. – Computational Linguistics and Intellectual Technologies, 2022. DOI: 10.28995/2075-7182-2022-21-114-131.

34. Krause, B., A. D. Gotmare, B. McCann, N. S. Keskar, S. Joty, R. Socher, N. F. Rajani. Gedi: Generative Discriminator Guided Sequence Generation. – arXiv Preprint arXiv:2009.06367. 2020. DOI: 10.18653/v1/2021.findings-emnlp.424.

35. Roiqoh, S., B. Zaman, K. Kartono. Analisis Sentimen Berbasis Aspek Ulasan Aplikasi Mobile JKN Dengan Lexicon Based dan Naïve Bayes. – Jurnal Media Informatika Budidarma, Vol. **7**, 2023, No 3, pp. 1582-1592.

Received: 02.05.2025. Revised version: 16.06.2025. Accepted: 26.06.2025