IJSA (2025), Vol. 9, No. 1, 100-116
DOI: https://doi.org/10.29244/ijsa.v9i1p100-116

Indonesian Journal of Statistics and Its Applications
Available at https://journal-stats.ipb.ac.id/index.php/ijsa
e-ISSN: 2599-0802

Exploring a Large Language Model on the ChatGPT Platform for Indonesian Text Preprocessing Tasks*

Cici Suhaeni[1‡], Sabrina Adnin Kamila[1], Fani Fahira[1], Muhammad Yusran[1], Gerry Alfa Dito[1]

[1]Study program on Statistics and Data Science, School of Data Science, Mathematics, and Informatics, IPB University, Indonesia
[‡]Corresponding author: cici_suhaeni@apps.ipb.ac.id

Abstract

Preprocessing is a crucial step in Natural Language Processing, especially for informal languages like Indonesian, which contain complex morphology, slang, abbreviations, and non-standard expressions. Traditional rule-based tools such as regex, IndoNLP, and Sastrawi are commonly used but often fall short in handling noisy, user-generated text. This study explores the capability of Large Language Model, particularly ChatGPT-o3, in performing Indonesian text preprocessing tasks, namely text cleaning, normalization, stopword removal, and stemming/lemmatization, and compares it to conventional rule-based approaches. Using two types of datasets, consisting of a small example dataset of five manually constructed sentences and a real-world dataset of 100 tweets about the Indonesian “Makan Bergizi Gratis” program, both preprocessing methods were applied and evaluated. Results show that ChatGPT-o3 performs equally well in text cleaning and significantly better in normalization. However, rule-based methods like IndoNLP and Sastrawi still outperform ChatGPT-o3 in stopword removal and stemming. These findings indicate that while ChatGPT-o3 demonstrates strong contextual understanding and linguistic flexibility, they may underperform in rigid, token-based operations without fine-tuning. This study provides initial insights into using Large Language Models as an alternative preprocessing engine for Indonesian text and highlights the need for hybrid approaches or improved prompt design in future applications.

Keywords: ChatGPT, indoNLP, Large Language Model, Sastrawi, Text Preprocessing.

*Received: Jun 2025; Reviewed: Jun 2025; Published: Jun 2025

1. Introduction

Text preprocessing is a fundamental stage in natural language processing (NLP), as it transforms raw, unstructured input into a cleaner and more standardized format suitable for downstream tasks such as classification, sentiment analysis, summarization, and intent detection. For low-resource languages such as Bahasa Indonesia, the preprocessing phase becomes more complex due to the language's rich morphology, agglutinative nature, and frequent use of informal expressions, non-standard abbreviations, and typographical errors—especially on social media platforms (Hasanah et al., 2018; Nugraheni et al., 2024; T. Rahman et al., 2019; Rianto et al., 2021). This linguistic diversity presents a critical challenge for the development of robust NLP pipelines.

Traditional rule-based approaches, including regular expression (`re`) scripts, the IndoNLP library, and the Sastrawi stemmer, have long been used in Indonesian NLP tasks such as text classification, sentiment analysis, and document indexing (Purbolaksono et al., 2020; Rosid et al., 2020; Setiabudi et al., 2021). These tools rely on deterministic heuristics such as pattern-matching, fixed dictionaries, and handcrafted rules. However, several studies report that these approaches struggle with non-formal Indonesian text, particularly in over-stemming, under-stemming, and failure to handle informal tokens, leading to a decline in model accuracy (Lubis et al., 2023; Purbolaksono et al., 2020; T. Rahman et al., 2019; Rianto et al., 2021). While modifications and hybrid preprocessing pipelines have been proposed (e.g., idtext_normalizer by Nugraheni et al., 2024), the rule-based paradigm remains limited by its rigidity and inability to contextualize.

The recent emergence of Large Language Models (LLMs) like models in ChatGPT platform has shifted the NLP landscape toward context-aware, prompt-based approaches. Trained on massive multilingual corpora, ChatGPT is capable of executing various linguistic tasks, including grammar correction, normalization, annotation, summarization, and translation—all through natural language prompts (Blüthgen, 2025; Hamarashid et al., 2023). Its architecture, based on transformers and self-attention mechanisms, allows it to model semantic dependencies in a way that traditional rule-based tools cannot (Dong et al., 2023; Lai et al., 2023). Several recent studies have started to evaluate the performance of ChatGPT in various NLP tasks, showing promising results. For example, Belal et al., (2023) demonstrated that ChatGPT outperforms lexicon-based sentiment annotation tools by a wide margin, with improvements of up to 25%. R. A. Rahman & Suyanto (2024) employed ChatGPT for abstractive summarization in Bahasa Indonesia and achieved coherent results close to human expectations. In a study focused on annotation quality, Nasution & Onan (2024) showed that while human annotators remain superior in nuanced tasks, LLM-generated labels for sentiment and topic classification still yielded competitive precision and recall. Similarly, Julianto et al., (2023) found that ChatGPT could serve as a viable alternative for preprocessing, achieving accuracy close to traditional tools such as Rapidminer and outperforming Google Bard in sentiment analysis preprocessing with Decision Tree and Naïve Bayes classifiers.

While these studies highlight ChatGPT's potential across NLP tasks, most of them focus on high-level outcomes such as classification or annotation quality, rather than on the specific step of text preprocessing. Moreover, evaluations often involve English datasets or do not explicitly compare ChatGPT's performance with rule-based methods for Indonesian text. A few studies address the limitations of existing preprocessing methods in informal or noisy contexts (Lubis et al., 2023; Rianto et al., 2021; Setiabudi et al., 2021), yet systematic explorations of ChatGPT's capability to perform structured preprocessing (e.g., cleaning, normalization, stemming, lemmatization) in Bahasa Indonesia remain scarce.

In response to the identified gap, this study is conducted with the following objectives:

1. Exploring the capabilities of a Large Language Model (LLM), particularly ChatGPT o3, in performing preprocessing tasks for Indonesian text, including the stages of text cleaning, normalization, stopword removal, and stemming/lemmatization.
2. Comparing the preprocessing results generated by rule-based approaches (regex, IndoNLP, and Sastrawi) with those generated through prompt-based interaction with ChatGPT.

Through this exploration, the study contributes in two main ways: First, by providing a set of effective and adaptable prompts for preprocessing Indonesian text using ChatGPT models. Second, by offering a preliminary investigation that may serve as the foundation for future quantitative evaluations or the integration of LLM-based preprocessing into broader NLP pipelines.

2. Metodology

2.1 Data

This study utilizes two types of textual data: example data and real-world data. The example data consists of five self-constructed Indonesian-language sentences related to skincare product reviews. These examples were manually created by the researcher to serve as controlled input for initial prompt testing and comparative preprocessing demonstrations.

The real-world data used in this study consists of a single dataset containing 100 tweets related to public discourse on *Makan Bergizi Gratis (MBG)*, a national food program issue in Indonesia. These tweets were collected using the keyword *“makan bergizi gratis”* through a web scraping process conducted in June 2025. The dataset reflects informal, user-generated content typical of social media platforms and was selected to evaluate the effectiveness of preprocessing methods in handling unstructured Indonesian text.

The datasets were used as raw input for the comparative preprocessing experiments between rule-based and LLM-based approaches. The example data served to evaluate prompt behavior in a controlled scenario, while the real-world data provided insights into preprocessing performance on informal, user-generated content typical of social media platforms.

2.2 The Data Analysis

The data analysis in this study was conducted through a three-stage process, involving both example and real-world datasets. The three stages included: (1) preprocessing of example data, (2) preprocessing of real-world data, and (3) evaluation of preprocessing results from both datasets using two different approaches—rule-based and LLM-based (ChatGPT). Each stage was designed to systematically compare the performance of these approaches in handling Indonesian-language text. The overall analysis workflow is illustrated in the research flow diagram presented in Figure 1.

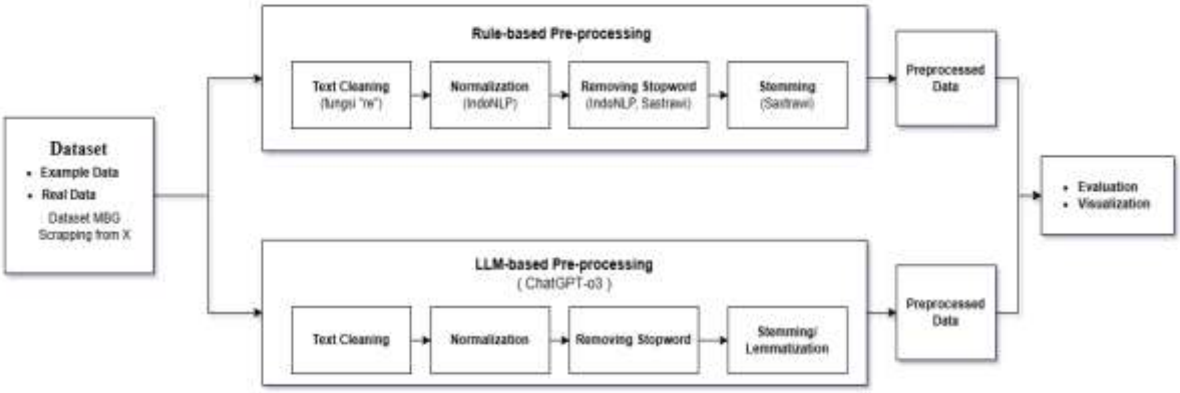


Figure 1: Research Workflow for Indonesian Text Preprocessing Using Rule-Based and LLM-Based Approaches

Stage 1: Preprocessing on Example Data

In this initial stage, a small dataset consisting of five manually created skincare review sentences was used to test and refine the preprocessing pipeline. Two approaches were implemented in parallel:

1. Rule-Based Approach, using:
    - Regular expression (`re`) functions in Python In this study, the `re` module—Python's built-in library for working with regular expressions—was used to perform basic text cleaning tasks such as removing URLs, punctuation, and extra whitespace. This module provides flexible functions for pattern matching and substitution in string preprocessing (Python, 2024)
    - IndoNLP library

      As part of the rule-based approach, this study incorporates selected functions from the IndoNLP preprocessing module developed by Hyuto
    - (Hyuto, n.d.) which is specifically designed for processing Indonesian text. While the module provides various utilities such as remove_html(), remove_url(), emoji_to_words(), and words_to_emoji(), this study utilizes only three core functions: remove_stopwords(),replace_slang(),and replace_word_elongation(). These functions are applied during the normalization and stopword removal stages to eliminate non-informative words, convert slang or informal abbreviations into their standard forms, and correct exaggerated character repetitions commonly found in informal Indonesian text.
    - Sastrawi stemmer

      In this study, the Sastrawi library was used as part of the rule-based preprocessing approach. Sastrawi is a popular open-source library in Python specifically designed for Indonesian stemming. It applies morphological rules to reduce words to their root form (e.g., *berjalan*, *berjalanlah*, *menjalankan* → *jalan*). In addition to stemming, Sastrawi also includes a built-in list of Indonesian stopwords, enabling basic stopword removal (Sastrawi, n.d.)

2. LLM-Based Approach, using ChatGPT-o3 via prompt-based interactions The LLM-based preprocessing in this study was performed using ChatGPT-o3 through prompt-based interactions. ChatGPT-o3 belongs to OpenAI's o-series models, which are designed with advanced reasoning capabilities through large-scale reinforcement learning on chains of thought (Openai, 2025). While not explicitly built for preprocessing, ChatGPT-o3 demonstrates strong contextual understanding, making it potentially capable of interpreting informal language, correcting spelling, and normalizing text—qualities that make it suitable for Indonesian text preprocessing tasks.

Both approaches applied the same preprocessing pipeline consisting of four sequential steps:

Step 1. Text Cleaning

The first step focused on removing irrelevant elements from the raw text, including URLs, HTML tags, punctuation, special characters, numbers, emojis, and extra whitespaces to ensure clean and consistent formatting.

Step 2. Text Normalization

Normalization aimed to standardize the linguistic form of the input text. All characters were first converted to lowercase to eliminate case sensitivity. Then, character elongations often used to express emotion or emphasis, such as baguuuusss, were corrected to their standard form (bagus). The process continued with spelling correction and translation of non-standard or abbreviated terms into formal Indonesian, for example, gk to tidak, and bgt to banget. Slang words and informal expressions were replaced with their formal equivalents, while any foreign language terms were translated into proper Bahasa Indonesia. Additionally, meaningless symbols or unreadable characters that did not contribute semantic value were removed.

Step 3. Stopword Removal

In this step, the text was filtered to remove common Indonesian stopwords, words that frequently appear in text but contribute little to the meaning of a sentence, such as *yang*, *dan*, or *itu*. The removal process was done carefully to ensure that the main message or semantic core of each sentence remained intact, allowing downstream NLP tasks to focus on more informative and content-rich terms.

Step 4. Stemming / Lemmatization

The final step involved stemming or lemmatization, where each word was reduced to its root or base form. This process was guided by the grammatical structure of Bahasa Indonesia and aimed to consolidate word variants into a single canonical form. For instance, words like *berjalan*, *berjalanlah*, and *menjalankan* would all be reduced to the base form *jalan*. This step helps in minimizing vocabulary size and improving the generalization ability of text-based models during further analysis.

Stage 2: Preprocessing on Real-World Data

The second stage replicated the full preprocessing pipeline from Stage 1, now applied to the real-world dataset, the twitter dataset related to the keyword *"makan bergizi gratis"*

Stage 3: Evaluation

The final stage of this study involved a combination of qualitative and semi-quantitative evaluations to assess the effectiveness of both rule-based and LLM-based preprocessing approaches. For the example dataset, a qualitative descriptive analysis was conducted by comparing the output of each preprocessing step. This comparison was presented in a tabular format, showcasing side-by-side results from the rule-based and LLM-based approaches to highlight specific differences and patterns.

For the real-world dataset, the evaluation was conducted in two ways. First, a quantitative analysis was performed by calculating the success rate of each preprocessing method in handling specific tasks. Each step—such as removing HTML tags, emojis, punctuation, and correcting elongated words—was evaluated independently. The success percentage for each step was computed using the following formula:

$$Success\ Rate\ (\%) = \left(\frac{Number\ of\ Successful\ Cases}{Total\ Relevant\ Cases}\right) \times 100$$

In this evaluation, *successful cases* refer to instances where a method correctly performed the intended preprocessing operation, and *total relevant cases* refer to the number of data entries requiring that specific preprocessing step, as defined in the stages 1 step 1 of this study. *Second*, a visual analysis was conducted using word cloud representations. Word clouds serve as a descriptive tool to illustrate the textual outcome of each preprocessing method. By visualizing the most frequent and prominent words retained after preprocessing, the word clouds help convey the overall effectiveness and cleanliness of the processed text. This allows for a general comparison of how well each method, rule-based or LLM-based—preserved meaningful content while eliminating noise, offering an intuitive understanding of preprocessing performance.

3. Results and Discussion

3.1 Example Dataset: Preprocessing Outcomes

The preprocessing results of the example dataset, consisting of five data entries, are presented in a descriptive qualitative format by displaying the processed text at each preprocessing stage. This aims to illustrate the transformation process from the original input to the final preprocessed output.

Table 1 presents the comparison of preprocessing outcomes for the text cleaning stage using both rule-based and LLM-based (ChatGPT-o3) approaches. The results demonstrate that all text cleaning components were successfully handled by both methods. This includes the removal of URLs, HTML tags, punctuation and special characters, numbers, emojis, and extra whitespaces.

The fact that both approaches consistently cleaned all elements across the example dataset suggests that rule-based and LLM-based methods exhibit equally effective capabilities in performing basic text cleaning operations. This indicates a strong baseline performance of ChatGPT-o3, matching that of well-established rule-based scripts in processing structurally noisy text data.

Table 1: Cleaning results on example data

Approach		Cleaning Result
Rule-based Pre-processing	*re*-Regex Function	1. aku sukaaa bangetttt produk ini cek di langsung yah recommended 2. pdkt ama serum ini dan ternyata baguuuss bgt sihhh kulitku jd kinclong parahhhh glowingmax

Approach		Cleaning Result
		3. gak ngerti lg ini tuh skincare yg terbaikkk efeknyaa cepet bngt hasilnya gajls klo ga rutin 4. serum ini tuh murce bgttt cocok bwt kulitku yg sensitif bgt jerawatku juga ilang loh hehe 5. itemnya woww deh sukaaaaaaaa bangettt padahal awalnya ragu tp hasilnya mantulllll skrg repeat order terussss
LLM-based Pre-processing	ChatGPT-o3	1. aku suka banget produk ini cek di langsung yah recommended 2. pdkt ama serum ini dan ternyata bagus banget sih kulitku jadi kinclong parah glowingmax 3. gak ngerti lg ini tuh skincare yg terbaik efeknya cepat bngt hasilnya gajls klo ga rutin 4. serum ini tuh murce bgttt cocok bwt kulitku yg sensitif bgt jerawatku juga ilang loh hehe 5. itemnya woww deh suka banget padahal awalnya ragu tp hasilnya mantulllll skrg repeat order terus

As presented in Table 2, the normalization results show a clear distinction between the two approaches. The LLM-based approach using ChatGPT-o3 demonstrates a more effective performance in handling various normalization aspects. ChatGPT-o3 was able to accurately transform slang words, abbreviations, and informal expressions into their formal Indonesian equivalents. The resulting sentences are more grammatically correct and appropriate for formal contexts, indicating the model's contextual understanding and semantic refinement capabilities.

In contrast, the rule-based approach failed to fully normalize several elements. Although it successfully applied lowercasing, it did not adequately convert slang or informal abbreviations, resulting in outputs that still contain non-standard words and casual language forms. This highlights the limitations of conventional tools when dealing with informal and varied linguistic structures typical in user-generated content. Overall, the findings suggest that the LLM-based approach offers a distinct advantage in text normalization, particularly in producing cleaner, more standardized, and contextually appropriate Indonesian sentences.

Table 2: Normalization results on example data

Approach		Normalization Result
Rule-based Pre-processing	IndoNLP	1. aku suka banget produk ini cek di langsung ya recommended 2. pdkt sama serum ini dan ternyata bagus banget sih kulitku jadi kinclong parah glowingmax 3. enggak mengerti lagi ini tuh skincare yang terbaik efeknya cepat banget hasilnya gajls kalo enggak rutin

Approach		Normalization Result
		4. serum ini tuh murce banget cocok buat kulitku yang sensitif banget jerawatku juga hilang loh hehe 5. itemnya wow deh suka banget padahal awalnya ragu tapi hasilnya mantul sekarang repeat order terus
LLM-based Pre-processing	ChatGPT-o3	1. saya sangat menyukai produk ini silakan cek langsung direkomendasikan 2. saya mencoba serum ini dan ternyata sangat bagus kulit saya menjadi sangat berkilau 3. saya tidak mengerti lagi ini adalah produk perawatan kulit yang terbaik efeknya sangat cepat namun hasilnya tidak jelas jika tidak rutin 4. serum ini murah dan sangat cocok untuk kulit saya yang sangat sensitif jerawat saya juga hilang 5. produk ini sangat baik saya sangat suka padahal awalnya ragu tetapi hasilnya sangat memuaskan sekarang saya memesan ulang terus menerus

Table 3 presents a comparison of stopword removal outcomes using three different methods: IndoNLP, Sastrawi, and ChatGPT-o3. Based on the results from the five example sentences, none of the methods were able to remove stopwords perfectly. However, a comparative observation reveals that ChatGPT-o3 performed slightly better than the other two, with Sastrawi showing better performance than IndoNLP.

Several specific findings support this conclusion. For instance, the word "ya" still appears in the IndoNLP output, while it is successfully removed in both Sastrawi and ChatGPT-o3 outputs. Similarly, informal interjections like "loh" and "wow deh" remain in the outputs of IndoNLP and Sastrawi, but are absent in the ChatGPT-o3 version. This indicates a more refined understanding of stopword relevance in ChatGPT-o3's output, even though the differences are subtle. While these conclusions are drawn from a limited set of short sentences, they suggest that ChatGPT-o3 holds a slight better in stopword removal. However, further validation using real-world data with more complex and noisy sentence structures is necessary. It is also worth noting that ChatGPT-o3's occasional retention of certain stopwords may be intentional—stemming from its design to preserve sentence meaning and coherence, rather than simply removing non-informative words without regard to context.

Table 3: Removing stopword results on example data

Approach		Removing Stopword Result
Rule-based Pre-processing	IndoNLP	1. suka banget produk cek langsung ya recomended 2. pdkt serum bagus banget sih kulitku kinclong parah glowingmax

Approach		Removing Stopword Result
		3. mengerti tuh skincare terbaik efeknya cepat banget hasilnya gajls kalo rutin 4. serum tuh murce banget cocok kulitku sensitif banget jerawatku hilang loh hehe 5. itemnya wow deh suka banget ragu hasilnya mantul repeat order
	Sastrawi	1. aku suka banget produk cek langsung recomended 2. pdkt sama serum dan ternyata bagus banget sih kulitku jadi kinclong parah glowingmax 3. enggak mengerti ini tuh skincare terbaik efeknya cepat banget hasilnya gajls kalo enggak rutin 4. serum tuh murce banget cocok buat kulitku sensitif banget jerawatku hilang loh hehe 5. itemnya wow deh suka banget padahal awalnya ragu hasilnya mantul sekarang repeat order terus
LLM-based Pre-processing	ChatGPT-o3	1. menyukai produk cek langsung direkomendasikan 2. mencoba serum bagus kulit berkilau 3. tidak mengerti produk perawatan kulit terbaik efeknya cepat hasilnya tidak jelas jika tidak rutin 4. serum murah cocok kulit sensitif jerawat hilang 5. produk baik suka ragu hasilnya memuaskan memesan ulang terus menerus

Based on the five example sentences, both the Sastrawi stemmer and ChatGPT-o3 appear to perform well in reducing words to their root or base forms. The outputs show that both approaches are capable of executing the stemming or lemmatization task appropriately. However, this observation is limited to the small and relatively simple dataset used in this example.

Because the sample data does not include a wide variety of affixed or morphologically complex words, it may not fully represent the challenges typically encountered in Indonesian stemming. Therefore, a more comprehensive analysis using real-world datasets is needed to better assess the consistency and reliability of both approaches under more diverse linguistic conditions. A more in-depth discussion of this evaluation is presented in Section 3.2.

Table 4: Stemming results on example data

Approach		Stemming Result
Rule-based Pre-processing	Sastrawi	1. aku suka banget produk cek langsung recommended

Approach		Stemming Result
		2. pdkt sama serum dan nyata bagus banget sih kulit jadi kinclong parah glowingmax 3. enggak erti ini tuh skincare baik efek cepat banget hasil gajls kalo enggak rutin 4. serum tuh murce banget cocok buat kulit sensitif banget jerawat hilang loh hehe 5. item wow deh suka banget padahal awal ragu hasil mantul sekarang repeat order terus
LLM-based Pre-processing	ChatGPT-o3	1. suka produk cek langsung rekomendasi 2. coba serum bagus kulit kilau 3. tidak erti produk rawat kulit baik efek cepat hasil tidak jelas jika tidak rutin 4. serum murah cocok kulit sensitif jerawat hilang 5. produk baik suka ragu hasil puas pesan ulang terus terus

3.2 Real Dataset: Preprocessing Outcomes

The preprocessing results of the real-world dataset were evaluated using two complementary approaches: a quantitative descriptive analysis and a visual representation through wordclouds. This section presents the insights gained from both evaluations to compare the performance of the rule-based and LLM-based (ChatGPT-o3) methods when applied to actual Indonesian user-generated text.

To conduct the LLM-based preprocessing, the model ChatGPT-o3 was prompted through a structured and detailed instruction using a prompt-based interaction format. The full prompt used was as follows:

> *"I have approximately 100 short texts written in Bahasa Indonesia.*
> *Please use your capabilities as an LLM (ChatGPT-o3) to perform text preprocessing, which includes Text Cleaning, Normalization,Stopword Removal, and Stemming/Lemmatization.*
> *The detailed steps for each preprocessing stage are as follows.*
> *1. Text Cleaning*
> - *Remove URLs*
> - *Remove HTML tags*
> - *Remove punctuation and special characters*
> - *Remove numbers*
> - *Handle emojis (remove or replace with appropriate text)*
> - *Remove extra whitespaces*
>
> *2. Normalization*
> - *Convert all letters to lowercase*
> - *Correct character elongation (e.g., "baguuuusss" → "bagus")*
> - *Fix spelling errors*
> - *Correct non-standard words (e.g., "gk" → "tidak", "bgt" → "banget")*
> - *Standardize common abbreviations (e.g., "pdkt" → "pendekatan")*
> - *Convert slang or informal words to standard formal words (e.g., "banget" → "sangat")*
> - *Translate or convert any non-Indonesian words or phrases into proper Bahasa Indonesia*
> - *Remove any meaningless characters or unreadable/non-linguistic words that do not add value to the sentence*
>
> *3. Stopword Removal*

- *Remove common and non-informative Indonesian stopwords*
- *Keep the main meaning of the sentence intact*

4. Stemming / Lemmatization

- *Convert each word into its root or base form using proper Bahasa Indonesia grammar*
- *Convert all affixed words into their root or base form in Bahasa Indonesia. The final output of this preprocessing stage should consist of standard and base words in proper Bahasa Indonesia.*

Important constraints:

- *Do not add any punctuation marks (e.g., commas, periods, exclamation marks, etc.)*
- *Do not rephrase or elaborate the sentence beyond what is necessary for proper normalization*
- *You may rearrange word order only if required to preserve meaning or grammar (e.g., "pengen banget" → "sangat ingin")*
- *Return only the final processed sentence in plain text*
- *Do not include any explanation or formatting*

Please return the result in this format: No | preprocessed text. I will provide the text one by one in this chat, using the format: No and the sentence. Please only process the sentence — the number is for ID purposes only. If you understand this instruction, I will begin sending the data."

1) Descriptive Quantitative Analysis

Table 5 summarizes the success rates of each preprocessing step across both methods. The success rate was calculated based on the proportion of cases correctly handled by each approach for each specific task. The findings are as follows:

- Text Cleaning: Both the rule-based and LLM-based methods achieved a perfect success rate of 100%, confirming that both techniques are equally effective at removing URLs, HTML tags, punctuation, numbers, emojis, and extra whitespaces. This result aligns with the observations from the example dataset.
- Normalization: ChatGPT-o3 significantly outperformed the rule-based approach (represented by IndoNLP) in this stage, achieving a success rate of 91.79%, compared to 72% for IndoNLP. This reinforces ChatGPT's strength in interpreting and transforming informal or non-standard expressions into formal Indonesian.
- Stopword Removal: The rule-based method using IndoNLP yielded the highest performance, followed closely by Sastrawi and then ChatGPT-o3, with only a slight margin of around 1% separating Sastrawi and GPT-o3.
- Stemming/Lemmatization: In this final stage, Sastrawi demonstrated a clear advantage, with an accuracy of 96%, while ChatGPT-o3 lagged behind at only 50%.

Table 5: Comparison of Success Rate of Text Preprocessing on Real-World Data

Preprocessing Step	Rule-based (%)			ChatGPT-o3
	re function	indoNLP	sastrawi	
Text Cleaning	100%			100%
Text Normalization		72.00%		91.79%
Stopword Removal		82.41%	60.67%	59.00%
Stemming			96.68%	50.00%

These findings confirm and expand upon the patterns identified in the example dataset. While both approaches handle text cleaning effectively, ChatGPT-o3 excels in normalization, suggesting its strength in understanding context and rephrasing informal expressions. Conversely, the rule-based tools outperform ChatGPT-o3 in stopword removal and stemming, likely due to the strict linguistic rules and dictionaries they rely on. This performance gap may be attributed to ChatGPT's behavior as a generative language model that tends to preserve sentence meaning. In some cases, it chooses to retain stopwords or affixed words if removing them might alter the semantic intent of the sentence. Another possible explanation lies in the prompt design, which may need further refinement to instruct the model more precisely.

Additionally, it is worth noting that in practical NLP implementations, not all tasks require aggressive stopword removal or stemming. Depending on the downstream application, such as sentiment analysis or text generation, preserving certain grammatical structures might even be beneficial.

Table 6: Example of preprocessed data using Rule-based and LLM-based methods

No	Text	Rule-based preprocesed text	LLM-based preprocessed text
1	Program Makan Bergizi Gratis (MBG) telah dimulai sebagai langkah awal untuk mewujudkan Generasi Emas Indonesia 2045. Program ini bertujuan untuk meningkatkan kualitas gizi anak-anak sejak dini. Lanjutkan MBG #BGNSiapkanGenerasiEmas rileks ala artis Wahabi plossa Trumper Khamenei https://t.co/BojL2m8iCQ	program makan gizi gratis mbak mulai langkah awal wujud generasi emas indonesia program tuju tingkat kualitas gizi anakanak sejak dini lanjut mbak bgnsiapkangenerasiemas rileks ala artis wahabi plossa trumper khamenei	program makan gizi gratis mbg mulai langkah awal wujud generasi emas indonesia kualitas gizi anak dini lanjut mbg rileks artis wahabi plossa trumper khamenei
2	@idhofairy @T4atak @The_EnzBi @MemeComicIndo Ngawur banget! Siapa yang tega kasih anaknya racun? Makan Bergizi Gratis jelas yang terbaik bikin anak sehat dan cerdas! Lanjutkan MBG https://t.co/EUGVOSydK2	idhofairy tatak theenzbi memecomicindo ngawur banget siapa tega kasih anak racun makan gizi gratis jelas baik bikin anak sehat cerdas lanjut mbak	ngawur sangat tega kasih anak racun makan gizi gratis jelas terbaik buat anak sehat cerdas lanjut mbg
3	Grab Indonesia mengalokasikan anggaran sebesar 1 juta dollar Amerika Serikat setara dengan Rp16 3 miliar untuk mendukung program Makan Bergizi Gratis (MBG) yang digencarkan oleh pemerintah. https://t.co/qm7axSI43W	grab indonesia alokasi anggar besar juta dollar amerika serikat tara rp miliar dukung program makan gizi gratis mbak gencar perintah	grab indonesia alok anggaran besar juta dollar amerika serikat setara rp miliar dukung program makan gizi gratis mbg gencar pemerintah

Table 6 presents examples of the preprocessing results that support the findings shown in Table 5. These examples illustrate the differences in outcomes between the rule-based and Large Language Model (LLM)-based approaches. In the cleaning stage for structured noise such as URLs, numbers, and punctuation, both methods demonstrate comparable effectiveness, with both the *re* function and ChatGPT-o3 achieving a perfect 100% success rate. However, a significant divergence in performance is observed in the normalization and stemming stages.

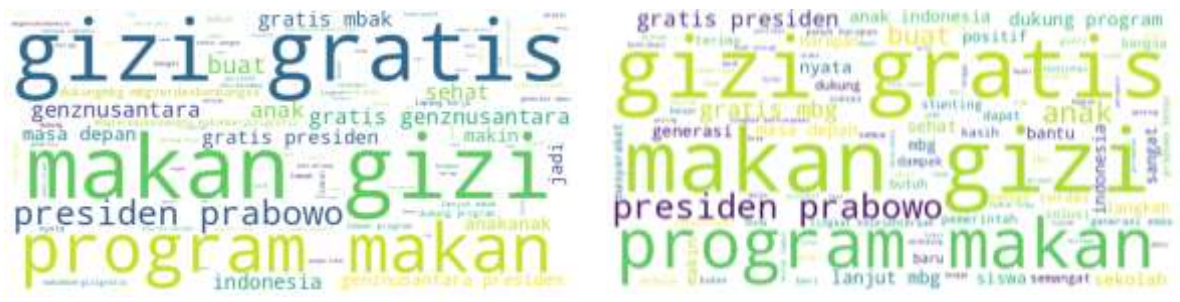
The rule-based approach excels in tasks that strictly apply linguistic rules. According to Table 5, in the Stemming process, the Sastrawi library achieves a superior accuracy rate of 96.68%. This advantage is qualitatively demonstrated by its ability to consistently reduce affixed words to their base forms, such as mewujudkan (to realize) to wujud (form) and pemerintah (government) to perintah (order). However, its main weakness lies in contextual understanding, which is reflected in its Text Normalization accuracy rate of only 72.00%. The reliance on a rigid dictionary leads to misinterpretations, for instance, with the acronym "MBG" being incorrectly normalized to "mbak," and the failure to normalize informal words like "banget".

Conversely, the LLM-based approach demonstrates clear superiority in tasks that demand semantic and contextual understanding. This is evidenced by its 91.79% success rate in Text Normalization, far surpassing the rule-based method. This figure confirms its ability to understand context to correctly interpret acronyms like "MBG" and normalize informal expressions such as "ngawur banget" to "ngawur sangat." Nevertheless, the LLM's performance is lower in more mechanical tasks. In the Stemming stage, its accuracy is only 50.00%, a stark contrast to Sastrawi, which is seen in its tendency to retain some words like pemerintah without reducing them. Its performance in Stopword Removal is also lower (59.00%), indicating that the model tends to prioritize semantic coherence over the strict application of linguistic rules.

In conclusion, this analysis reveals a clear trade-off between the linguistic precision of the rule-based approach and the semantic understanding of the LLM-based approach. The choice of a preprocessing method must be a strategic decision: the rule-based approach excels for tasks demanding high morphological integrity, while the LLM offers significant advantages for applications that are sensitive to meaning and context.

2) Visual Analysis

The results of the preprocessing visualization using wordclouds are presented in Figure 2, which compares outputs from both approaches: (a) the rule-based method and (b) the LLM-based method (ChatGPT-o3). Briefly, the two wordclouds appear relatively similar in terms of dominant words and overall text distribution. However, upon closer inspection, a notable discrepancy emerges. Specifically, the word “mbg,” which is the core topic of the dataset, remains clearly visible in the LLM-based wordcloud, whereas in the rule-based version, it is replaced by the unrelated word “mbak.” This significant error originates from the normalization step performed by the rule-based approach, particularly due to the function `replace_slang()` from the IndoNLP library, which incorrectly transforms "mbg" into "mbak." This mistake alters the semantic representation of the data and highlights a critical limitation in rule-based systems when handling domain-specific abbreviations or contextual slang.



(a) (b)

Figure 2: Word Clouds of Text Preprocessing Results: (a) Rule-Based and (b) LLM-Based

This observation visually reinforces the findings from the descriptive analysis in Table 5, where the LLM-based approach outperformed the rule-based method in the normalization stage. It underscores a broader challenge in Indonesian NLP: the lack of comprehensive and context-aware slang or abbreviation dictionaries within existing preprocessing tools. As a result, rule-based approaches may introduce errors when attempting to standardize informal expressions, while LLMs such as ChatGPT-o3 demonstrate better contextual understanding, preserving the original meaning more accurately during normalization.

4. Conclusion and Recommendation

This study explored the capabilities of Large Language Models (LLMs), particularly ChatGPT-o3, in performing text preprocessing for Indonesian-language text, focusing on a social media X dataset. The preprocessing process involved four main stages: text cleaning, normalization, stopword removal, and stemming/lemmatization. Two types of datasets were used: an example dataset of five self-constructed sentences and a real-world dataset consisting of 100 tweets about Makan Bergizi Gratis (MBG) program. In the example dataset, ChatGPT-o3 relatively outperformed the rule-based approach across all preprocessing stages. This may be attributed to the short and relatively noise-free structure of the sample texts, which allowed the model to apply transformations more accurately and consistently.

On the real-world dataset, both methods succeeded in the text cleaning stage. However, performance varied in the other stages: ChatGPT-o3 clearly outperformed the rule-based method in normalization, IndoNLP performed better in stopword removal, and Sastrawi was more effective in stemming. These results indicate that ChatGPT-o3, as a language model, tends to preserve semantic integrity, which enhances its normalization performance but may hinder its ability to perform rigid linguistic simplification such as stopword removal and stemming. Overall, while LLM-based preprocessing offers considerable advantages in understanding and transforming natural language, combining it with rule-based techniques may provide a more robust and balanced solution, especially when dealing with noisy, informal user-generated content.

This study is limited by the use of only one real-world dataset and a single LLM model due to platform constraints. Future studies should explore preprocessing performance using larger and more diverse datasets, experiment with optimized prompting strategies, and leverage OpenAI API calls for improved scalability. Additionally, future work could investigate the impact of fine-tuning LLMs on

Indonesian-specific preprocessing tasks, which may significantly improve their accuracy in structured linguistic operations such as stopword removal and stemming.

References

Belal, M., She, J., & Wong, S. (2023). *Leveraging ChatGPT As Text Annotation Tool For Sentiment Analysis*. https://arxiv.org/pdf/2306.17177

Blüthgen, C. (2025). Technical foundations of large language models[Technische Grundlagen großer Sprachmodelle]. *Radiologie*, *65*(4), 227–234. https://doi.org/10.1007/s00117-025-01427-z

Dong, Y., Xiao, C., & Oyamada, M. (2024). *Large Language Models as Data Preprocessors*. 3–6.

Hamarashid, H. K., Karim, L. T., & Muhammed, D. A. (2023). ChatGPT and Large Language Models: Unraveling Multifaceted Applications, Hallucinations, and Knowledge Extraction. *Indonesian Journal of Curriculum and Educational Technology Studies*, *11*(2), 60–70. https://doi.org/10.15294/IJCETS.V11I2.75617

Hasanah, U., Astuti, T., Wahyudi, R., Rifai, Z., & Pambudi, R. A. (2018). An experimental study of text preprocessing techniques for automatic short answer grading in Indonesian. *Proceedings - 2018 3rd International Conference on Information Technology, Information Systems and Electrical Engineering, ICITISEE 2018*, 230–234. https://doi.org/10.1109/ICITISEE.2018.8720957

Hyuto. (n.d.). *IndoNLP*. https://hyuto.github.io/indo-nlp/

Julianto, I. T., Kurniadi, D., & Jr, B. B. B. (2023). ENHANCING SENTIMENT ANALYSIS WITH CHATBOTS: A COMPARATIVE STUDY OF TEXT PRE-PROCESSING. *Jurnal Teknik Informatika (Jutif)*, *4*(6), 1419–1430. https://doi.org/10.52436/1.JUTIF.2023.4.6.1448

Lai, V. D., Ngo, N. T., Veyseh, A. P. Ben, Man, H., Dernoncourt, F., Bui, T., & Nguyen, T. H. (2023). ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 13171–13189. https://doi.org/10.18653/v1/2023.findings-emnlp.878

Lubis, A. R., Lase, Y. Y., Rahman, D. A., & Witarsyah, D. (2023). Improving Spell Checker Performance for Bahasa Indonesia Using Text Preprocessing Techniques with Deep Learning Models. *Ingenierie Des Systemes d'Information*, *28*(5), 1335–1342. https://doi.org/10.18280/ISI.280522

Nasution, A. H., & Onan, A. (2024). ChatGPT Label: Comparing the Quality of Human-Generated and LLM-Generated Annotations in Low-Resource Language NLP Tasks. *IEEE Access*, *12*(April), 71876–71900. https://doi.org/10.1109/ACCESS.2024.3402809

Nugraheni, E., Haekal, F. I., Arisal, A., & Perdana, R. S. (2024). Optimizing Indonesian Tweet Preprocessing on Halal Domain. *International Conference on Computer, Control, Informatics and Its Applications, IC3INA*, *2024*, 434–439. https://doi.org/10.1109/IC3INA64086.2024.10732128

Openai, T. (2025). *OpenAI o3 and o4-mini System Card*. 1–33.

Purbolaksono, M. D., Reskyadita, F. D., Adiwijaya, Suryani, A. A., & Huda, A. F. (2020). Indonesian Text Classification using Back Propagation and Sastrawi Stemming Analysis with Information Gain for Selection Feature. *International Journal on Advanced Science, Engineering and Information Technology*, *10*(1), 234–238. https://doi.org/10.18517/IJASEIT.10.1.8858

Python, S. (2024). *re — Regular expression operations*. https://docs.python.org/3/library/re.html

Rahman, R. A., & Suyanto. (2024). Performance Analysis of ChatGPT for Indonesian Abstractive Text Summarization. *Proceedings - International Seminar on Intelligent Technology and Its Applications, ISITIA*, *2024*, 477–482. https://doi.org/10.1109/ISITIA63062.2024.10668361

Rahman, T., Agustin, F. E. M., & Rozy, N. F. (2019). Normalization of Unstructured Indonesian Tweet Text for Presidential Candidates Sentiment Analysis. *2019 7th International Conference on Cyber and IT Service Management, CITSM 2019*, 2019. https://doi.org/10.1109/CITSM47753.2019.8965324

Rianto, Mutiara, A. B., Wibowo, E. P., & Santosa, P. I. (2021). Improving the accuracy of text classification using stemming method, a case of non-formal Indonesian conversation. *Journal of Big Data*, *8*(1), 1–16. https://doi.org/10.1186/S40537-021-00413-1/FIGURES/6

Rosid, M. A., Fitrani, A. S., Astutik, I. R. I., Mulloh, N. I., & Gozali, H. A. (2020). Improving Text Preprocessing for Student Complaint Document Classification Using Sastrawi. *IOP Conference Series: Materials Science and Engineering*, *874*(1), 012017. https://doi.org/10.1088/1757-899X/874/1/012017

Sastrawi. (n.d.). *Sastrawi*. ttps://github.com/sastrawi/sastrawi

Setiabudi, R., Iswari, N. M. S., & Rusli, A. (2021). Enhancing text classification performance by preprocessing misspelled words in Indonesian language. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, *19*(4), 1234–1241. https://doi.org/10.12928/TELKOMNIKA.V19I4.20369

Belal, M., She, J., & Wong, S. (2023). *Leveraging ChatGPT As Text Annotation Tool For Sentiment Analysis*. https://arxiv.org/pdf/2306.17177

Blüthgen, C. (2025). Technical foundations of large language models[Technische Grundlagen großer Sprachmodelle]. *Radiologie*, *65*(4), 227–234. https://doi.org/10.1007/s00117-025-01427-z

Dong, Y., Xiao, C., & Oyamada, M. (2024). *Large Language Models as Data Preprocessors*. 3–6.

Hamarashid, H. K., Karim, L. T., & Muhammed, D. A. (2023). ChatGPT and Large Language Models: Unraveling Multifaceted Applications, Hallucinations, and Knowledge Extraction. *Indonesian Journal of Curriculum and Educational Technology Studies*, *11*(2), 60–70. https://doi.org/10.15294/IJCETS.V11I2.75617

Hasanah, U., Astuti, T., Wahyudi, R., Rifai, Z., & Pambudi, R. A. (2018). An experimental study of text preprocessing techniques for automatic short answer grading in Indonesian. *Proceedings - 2018 3rd International Conference on Information Technology, Information Systems and Electrical Engineering, ICITISEE 2018*, 230–234. https://doi.org/10.1109/ICITISEE.2018.8720957

Hyuto. (n.d.). *IndoNLP*. https://hyuto.github.io/indo-nlp/

Julianto, I. T., Kurniadi, D., & Jr, B. B. B. (2023). ENHANCING SENTIMENT ANALYSIS WITH CHATBOTS: A COMPARATIVE STUDY OF TEXT PRE-PROCESSING. *Jurnal Teknik Informatika (Jutif)*, *4*(6), 1419–1430. https://doi.org/10.52436/1.JUTIF.2023.4.6.1448

Lai, V. D., Ngo, N. T., Veyseh, A. P. Ben, Man, H., Dernoncourt, F., Bui, T., & Nguyen, T. H. (2023). ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 13171–13189. https://doi.org/10.18653/v1/2023.findings-emnlp.878

Lubis, A. R., Lase, Y. Y., Rahman, D. A., & Witarsyah, D. (2023). Improving Spell Checker Performance for Bahasa Indonesia Using Text Preprocessing Techniques with Deep Learning Models. *Ingenierie Des Systemes d'Information*, *28*(5), 1335–1342. https://doi.org/10.18280/ISI.280522

Nasution, A. H., & Onan, A. (2024). ChatGPT Label: Comparing the Quality of Human-Generated and LLM-Generated Annotations in Low-Resource Language NLP Tasks. *IEEE Access*, *12*(April), 71876–71900. https://doi.org/10.1109/ACCESS.2024.3402809

Nugraheni, E., Haekal, F. I., Arisal, A., & Perdana, R. S. (2024). Optimizing Indonesian Tweet Preprocessing on Halal Domain. *International Conference on Computer, Control, Informatics and Its Applications, IC3INA*, *2024*, 434–439. https://doi.org/10.1109/IC3INA64086.2024.10732128

Openai, T. (2025). *OpenAI o3 and o4-mini System Card*. 1–33.

Purbolaksono, M. D., Reskyadita, F. D., Adiwijaya, Suryani, A. A., & Huda, A. F. (2020). Indonesian Text Classification using Back Propagation and Sastrawi Stemming Analysis with Information Gain for Selection Feature. *International Journal on Advanced Science, Engineering and Information Technology*, *10*(1), 234–238. https://doi.org/10.18517/IJASEIT.10.1.8858

Python, S. (2024). *re — Regular expression operations*. https://docs.python.org/3/library/re.html

Rahman, R. A., & Suyanto. (2024). Performance Analysis of ChatGPT for Indonesian Abstractive Text Summarization. *Proceedings - International Seminar on Intelligent Technology and Its Applications, ISITIA*, *2024*, 477–482. https://doi.org/10.1109/ISITIA63062.2024.10668361

Rahman, T., Agustin, F. E. M., & Rozy, N. F. (2019). Normalization of Unstructured Indonesian Tweet Text for Presidential Candidates Sentiment Analysis. *2019 7th International Conference on Cyber and IT Service Management, CITSM 2019*, 2019. https://doi.org/10.1109/CITSM47753.2019.8965324

Rianto, Mutiara, A. B., Wibowo, E. P., & Santosa, P. I. (2021). Improving the accuracy of text classification using stemming method, a case of non-formal Indonesian conversation. *Journal of Big Data*, *8*(1), 1–16. https://doi.org/10.1186/S40537-021-00413-1/FIGURES/6

Rosid, M. A., Fitrani, A. S., Astutik, I. R. I., Mulloh, N. I., & Gozali, H. A. (2020). Improving Text Preprocessing for Student Complaint Document Classification Using Sastrawi. *IOP Conference Series: Materials Science and Engineering*, *874*(1), 012017. https://doi.org/10.1088/1757-899X/874/1/012017

Sastrawi. (n.d.). *Sastrawi*. ttps://github.com/sastrawi/sastrawi

Setiabudi, R., Iswari, N. M. S., & Rusli, A. (2021). Enhancing text classification performance by preprocessing misspelled words in Indonesian language. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, *19*(4), 1234–1241. https://doi.org/10.12928/TELKOMNIKA.V19I4.20369