JCTA Journal of Computing Theories and Applications ISSN:3024-9104

Research Article

Quantifying the Impact of Text Preprocessing on IndoBERT Fine-Tuning for Indonesian Informal Culinary Sentiment Analysis

Rahmat Budianoor, Setyo Wahyu Saputro *, Friska Abadi, Radityo Adi Nugroho, and Andi Farmadi

Department of Computer Science, Lambung Mangkurat University, Banjarbaru 70714, South Kalimantan, Indonesia; e-mail : 2211016210032@mhs.ulm.ac.id; setyo.saputro@ulm.ac.id; friska.abadi@ulm.ac.id; radityo.adi@ulm.ac.id; andifarmadi@ulm.ac.id

* Corresponding Author : Setyo Wahyu Saputro

Abstract: Indonesian culinary comments on social media platforms such as Instagram are characterized by informal spelling, regional language mixing, slang expressions, and emojis, posing substantial challenges for automated sentiment classification. While IndoBERT has demonstrated strong performance across Indonesian natural language processing tasks, the contribution of individual preprocessing components to fine-tuning performance on informal text remains underexplored, particularly in the culinary domain. This study addresses this gap by conducting a systematic preprocessing ablation study on IndoBERT-Base fine-tuning for Indonesian culinary sentiment classification, accompanied by a comparative evaluation against Naive Bayes with TF-IDF, SVM with TF-IDF, and BiLSTM as representative baselines. A dataset of 3,500 manually labeled Instagram culinary comments across three sentiment classes was used, with a stratified 80/10/10 split. Six preprocessing variants were evaluated under identical experimental conditions to isolate the contribution of each component. The results show that slang normalization is the most impactful single preprocessing step, yielding a macro F1-score gain of +0.0609 over the no-preprocessing baseline, while the full pipeline achieves an accuracy of 0.8800 and a macro F1-score of 0.8465. IndoBERT-Base with the full pipeline outperforms all baselines across all evaluation metrics. Per-class analysis reveals that the negative class achieves the lowest F1-score of 0.7600, with sarcastic expressions and Banjar regional vocabulary identified as primary sources of misclassification. These findings indicate that preprocessing decisions have a measurable and non-uniform effect on IndoBERT fine-tuning performance. In this study, slang normalization provides the most substantial individual contribution in bridging the vocabulary gap between informal user-generated text and the model's pre-training distribution.

Keywords: Comparative sentiment classification; Fine-Tuning; IndoBERT; Indonesian language processing; Natural language processing; Preprocessing ablation sudy; Sentiment analysis; Social media analytics.

Received: April, 10th 2026
Revised: April, 28th 2026
Accepted: April, 30th 2026
Published: May, 8th 2026

1. Introduction

Online consumer reviews on social media have become a primary source of information in the culinary sector, as prospective customers increasingly rely on peer-generated opinions before making dining decisions. Prior hospitality research has demonstrated that textual sentiment extracted from online reviews is significantly associated with restaurant performance and profitability, providing insights beyond numerical ratings alone [1], [2]. In this context, Indonesia's culinary sector holds particular macroeconomic significance. The culinary subsector has been identified as a leading driver within Indonesia's creative economy, with strong capacity to stimulate economic activity from both the demand and supply sides [3]. At the SME level, culinary businesses constitute one of the most competitive subsectors in the national creative industry landscape, generating substantial employment and contributing meaningfully to regional and national economic output [4].

In Indonesia, culinary discussions are predominantly conducted through Instagram, where comment sections generate large volumes of user-generated feedback in highly informal language. Unlike conventional review corpora, Indonesian culinary comments on social media are typically short, noisy, and linguistically complex, frequently containing slang expressions, non-standard spelling, elongated characters, regional language usage, emojis, and code-mixing between Indonesian and local languages such as Banjar [5], [6]. The scale and variability of this feedback make manual inspection impractical, creating a clear need for automated sentiment classification [7], [8].

Earlier studies on Indonesian food and beverage sentiment analysis predominantly utilized conventional machine learning algorithms, including Naive Bayes and SVM, coupled with TF-IDF for feature representation [9]. While these methods demonstrated feasibility, they rely heavily on manual feature engineering and often struggle to capture contextual semantic dependencies in highly variable informal text [10]. To address these limitations, subsequent studies shifted toward deep learning architectures, with Bidirectional Long Short-Term Memory (BiLSTM) networks offering improved sequential context modeling over traditional classifiers [11], [12]. More recently, transformer-based pre-trained language models have significantly advanced Indonesian natural language processing. IndoBERT, pre-trained on large-scale Indonesian corpora [13], [14], has been successfully fine-tuned for sentiment classification across multiple domains, including e-commerce reviews [15], healthcare applications [16], aspect-based sentiment analysis [17], and travel user-generated content [18], consistently outperforming both traditional machine learning and recurrent neural network baselines.

Despite these advancements, two critical gaps remain. First, most existing IndoBERT-based sentiment studies treat preprocessing as a fixed, monolithic pipeline without empirically evaluating the individual contribution of each preprocessing component to model performance [19]. This limitation is particularly relevant for informal Indonesian text, where specific preprocessing decisions—such as slang normalization, emoji conversion, or elongation reduction—may have substantially different effects on fine-tuning outcomes [20], [21]. Prior studies have shown that preprocessing choices can significantly influence BERT-based classifiers on informal text [22], and that preprocessing variations can shift accuracy by up to 25% [19]; however, these studies do not specifically address the Indonesian context with its distinct linguistic characteristics. Second, to the best of the authors' knowledge, no published study has systematically applied IndoBERT fine-tuning with ablation-based preprocessing analysis to the Indonesian culinary domain, nor provided a controlled multi-baseline comparison spanning traditional machine learning, deep learning, and transformer paradigms within a unified experimental framework.

This study addresses these gaps by conducting a systematic preprocessing ablation study for IndoBERT-Base fine-tuning applied to Indonesian culinary Instagram comment sentiment classification, accompanied by a comparative evaluation against Naive Bayes with TF-IDF, SVM with TF-IDF, and BiLSTM as representative baselines. Six preprocessing variants are designed to isolate the contribution of individual components, including slang normalization, elongation reduction, and emoji-to-token conversion, in order to identify which steps most strongly influence fine-tuning performance on informal Indonesian text. A manually labeled dataset of 3,500 Instagram culinary comments across three sentiment classes (positive, neutral, negative) is used under consistent experimental conditions. The study further incorporates per-class performance analysis and qualitative error analysis to characterize model behavior under class imbalance and domain-specific linguistic challenges, including Banjar regional language expressions [6] and sarcasm [23].

The main contributions of this study are as follows. First, this study presents a systematic preprocessing ablation framework for IndoBERT fine-tuning on informal Indonesian text—an approach that, to the best of the authors' knowledge, has not been applied in the Indonesian culinary domain—quantifying the individual contribution of each preprocessing component and identifying slang normalization as the most impactful single step, with a macro F1-score gain of +0.0609 over the no-preprocessing baseline. Second, this study provides a controlled multi-paradigm comparison of IndoBERT-Base, SVM, BiLSTM, and Naive Bayes under identical preprocessing and evaluation conditions for Indonesian culinary sentiment classification, demonstrating that IndoBERT-Base achieves superior performance, with an accuracy of 0.8800 and a macro F1-score of 0.8465. Third, this study delivers per-class performance analysis and qualitative error analysis, revealing the role of class imbalance, Banjar

code-mixing, and sarcasm as key sources of misclassification, thereby providing actionable insights for future research on low-resource regional language adaptation in Indonesian sentiment analysis.

The remainder of this paper is organized as follows. Section 2 reviews related work on Indonesian sentiment analysis, preprocessing for informal text, IndoBERT fine-tuning, and class imbalance in natural language processing. Section 3 describes the proposed method, including dataset preparation, preprocessing variant design, model training configuration, and evaluation metrics. Section 4 presents and discusses the experimental results, including ablation outcomes, multi-baseline comparison, per-class analysis, and error analysis. Section 5 concludes the paper and outlines directions for future research.

2. Related Work

2.1. Sentiment Analysis for Indonesian Culinary and Food Reviews

Sentiment analysis has been established as a valuable tool for extracting business intelligence from online consumer feedback, particularly in the hospitality and culinary sectors. Studies such as [1], [2] demonstrate that textual sentiment from online reviews is significantly correlated with restaurant profitability, providing insights beyond numerical ratings alone. As the volume of digital feedback continues to grow, manual inspection becomes infeasible, thereby increasing the demand for automated sentiment classification systems [7], [8].

Within the Indonesian culinary context, the linguistic characteristics of social media comments present additional challenges for automated classification. These texts frequently contain slang, elongated characters, emojis, and code-mixing with regional languages such as Banjar and Javanese [5], [24]. Despite the growing use of IndoBERT in Indonesian NLP, studies specifically applying IndoBERT fine-tuning to the culinary domain remain limited. The informal nature of these comments suggests that preprocessing plays a critical role in shaping classification outcomes, making this domain particularly suitable for systematically investigating the contribution of individual preprocessing components.

2.2. IndoBERT and Transformer-Based Sentiment Analysis

IndoBERT was introduced through two foundational works: [14], which proposed IndoBERTweet, a pre-trained language model tailored for informal Indonesian text based on Twitter data, and [13], which released the IndoNLU benchmark, including the SmSA sentiment dataset and indobenchmark model variants. Both models adopt the BERT bidirectional transformer architecture [25], enabling context-sensitive representations that outperform static word embeddings, particularly for informal Indonesian text.

Subsequent studies have confirmed the effectiveness of IndoBERT fine-tuning across diverse domains, including e-commerce reviews [15], healthcare applications [16], aspect-based sentiment analysis [17], travel user-generated content [18], and general Indonesian benchmarks [26]. Table 1 summarizes these studies and reveals a consistent pattern: most IndoBERT-based sentiment models employ a fixed preprocessing pipeline without evaluating the contribution of individual preprocessing components. This suggests that the relationship between specific preprocessing decisions and fine-tuning performance on informal Indonesian text remains underexplored and warrants systematic investigation.

Table 1. Summary of IndoBERT fine-tuning studies for Indonesian sentiment analysis

Ref	Domain	Model Architecture	Preprocessing Strategy	Preprocessing Ablation
[15]	E-commerce	BERT + CNN + RNN	Fixed pipeline	No
[16]	Healthcare	IndoBERT	Fixed pipeline	No
[17]	Customer reviews	IndoBERT	Fixed pipeline	No
[18]	Travel UGC	IndoBERT	Fixed pipeline	No
[26]	General Indonesian	IndoBERT + BiLSTM	Fixed pipeline	No
Proposed	Culinary Instagram	IndoBERT-Base	Controlled ablation	Yes

2.3. Preprocessing for Transformer-Based Models and Ablation Studies

Preprocessing decisions for transformer-based models on informal text have received increasing attention in recent NLP literature. A comprehensive analysis in [19] shows that preprocessing can shift transformer accuracy by up to 25%, challenging the assumption that BERT-based models are inherently robust to noisy input. Multilingual evaluation in [22] further demonstrates that preprocessing effects vary across languages and text registers, with informal social media text being particularly sensitive to normalization strategies.

For Indonesian, [20] reports that slang normalization significantly improves downstream NLP performance, while [21] shows that selective emoji-to-text conversion is more effective than simple emoji removal for sentiment-bearing content. These findings indicate that preprocessing for IndoBERT fine-tuning on informal Indonesian text is not merely a preparatory step but a critical design variable requiring systematic evaluation.

Ablation studies have become a standard approach for isolating the contribution of individual components in NLP pipelines. However, the application of ablation design to preprocessing in transformer fine-tuning remains limited. As summarized in Table 1, most IndoBERT studies adopt a fixed preprocessing pipeline without measuring the marginal impact of each step [15]–[18]. A comparative deep learning study in [27] further shows that preprocessing choices—particularly slang normalization and emoji handling—directly influence classification outcomes across model architectures. These findings reinforce the need for systematic preprocessing evaluation rather than ad hoc pipeline design. The present study addresses this gap by applying a controlled one-component-at-a-time ablation design across six preprocessing variants.

2.4. Baseline Models for Evaluation Context

To provide a rigorous evaluation context, this study compares IndoBERT-Base against three representative baselines spanning distinct modeling paradigms: Naive Bayes with TF-IDF (traditional machine learning), SVM with TF-IDF (traditional machine learning), and BiLSTM (deep learning). This cross-paradigm design follows prior work [27], [28], which shows that evaluating transformer models alongside traditional and recurrent baselines under identical conditions yields more interpretable and generalizable insights.

For Indonesian social media data, previous studies have reported a consistent performance hierarchy across these paradigms [29], [30], providing a validated reference for contextualizing the results of the preprocessing ablation in this study. It should be noted that these models are included solely as baselines to support evaluation and are not the primary focus of the contribution.

2.5. Class Imbalance and Misclassification Challenges in Indonesian Sentiment Analysis

Class imbalance is a common issue in user-generated sentiment datasets, as online platforms tend to attract predominantly positive feedback. Study [31] shows that deep learning models systematically underperform on minority classes and recommends macro F1-score rather than accuracy as the primary evaluation metric for imbalanced datasets—an approach adopted in this study. This is further supported by [32], which demonstrates that accuracy can be misleadingly high despite poor minority-class performance, whereas macro F1 provides a more balanced evaluation. In the Indonesian context, [33] confirms that class imbalance significantly degrades minority-class performance and shows that SMOTE-based resampling can mitigate this effect.

Beyond class imbalance, two domain-specific challenges are particularly relevant to culinary Instagram comments. First, sarcasm and irony introduce systematic errors in sentiment classification. The IdSarcasm benchmark [23] shows that Indonesian PLMs, including IndoBERT variants, struggle to detect sarcasm without explicit markers, while [34] reports that sarcastic expressions are often misclassified as positive. Second, regional language mixing introduces out-of-vocabulary terms that are difficult to normalize during preprocessing. The NusaX benchmark [6] demonstrates that model performance degrades significantly on regional languages such as Banjar, confirming that regional vocabulary remains a persistent source of error.

2.6. Research Gap

Based on the reviewed literature, three main gaps motivate this study. First, although IndoBERT has been widely applied to Indonesian sentiment analysis across multiple domains, the individual contribution of preprocessing components to fine-tuning performance has not been systematically evaluated through controlled ablation, despite evidence that preprocessing can significantly influence transformer performance [19]. Second, IndoBERT fine-tuning has not been specifically explored in the Indonesian culinary domain, which presents unique linguistic challenges, including Banjar code-mixing, sarcasm, and extensive emoji usage. Third, no prior study provides a unified cross-paradigm comparison of traditional machine learning, deep learning, and transformer models under identical experimental conditions for this domain. To address these gaps, this study conducts a preprocessing ablation analysis combined with a multi-baseline comparison, supported by per-class evaluation and qualitative error analysis.

To guide the experimental design, this study is structured around three research objectives. The first objective is to identify which preprocessing component contributes most to IndoBERT-Base fine-tuning performance for Indonesian culinary Instagram sentiment classification. The second objective is to evaluate how IndoBERT-Base with a full preprocessing pipeline compares against traditional machine learning and deep learning baselines under identical conditions. The third objective is to examine the effect of class imbalance on per-class performance and to identify the most challenging sentiment category.

3. Proposed Method

This study employs a chronological experimental pipeline consisting of five stages: dataset preparation and labeling, preprocessing ablation design, model training and fine-tuning, comparative evaluation, and per-class and error analysis. Figure 1 illustrates the overall research workflow.

3.1. Dataset Collection and Labeling

The dataset was obtained by scraping publicly accessible comments from Instagram using the ExportComments tool (exportcomments.com), a web-based comment extraction service that retrieves publicly visible comment data from social media posts. Data collection was conducted from May to August 2025, covering a four-month period to capture a sufficient volume of naturalistic culinary discourse.

The target accounts were selected based on two criteria. First, active Instagram accounts belonging to small and medium culinary businesses (UMKM kuliner) located in South Kalimantan were included, defined as accounts that posted culinary content at least once per month during the collection period. Second, accounts belonging to culinary influencers who had been endorsed by or who actively promoted South Kalimantan UMKM culinary products were also included, as their comment sections reflect genuine consumer opinions about the promoted products. This selection strategy ensures that the collected comments are grounded in authentic South Kalimantan culinary discourse and reflect informal linguistic patterns, including regional Banjar expressions, which are central to this study [9].

Comments were included if they were written primarily in Indonesian or a recognizable regional language variant and contained at least one discernible sentiment signal. Comments were excluded if they were duplicates, contained only emojis without accompanying text, were off-topic or spam, or could not be processed as valid Indonesian text. After this screening stage, the final dataset comprised 3,500 comments distributed across three sentiment classes: Positive (2,061 samples, 58.9%), Neutral (932 samples, 26.6%), and Negative (507 samples, 14.5%). This positive-skewed distribution reflects patterns commonly observed in culinary social media datasets and constitutes a class imbalance, whose implications are discussed in Section 4.

Labels were assigned through a structured manual annotation process guided by a written annotation protocol. To support consistent initial judgments, a lexicon-based reference was used as a preliminary aid, drawing on the InSet Lexicon for Indonesian text alongside domain-relevant culinary and Banjar vocabulary. However, the final ground-truth label for each comment was determined through human judgment rather than automated scoring, to ensure label quality for informal and contextually nuanced text. Each comment was labeled Positive if it expressed satisfaction, praise, or recommendation; Negative if it expressed

complaint, disappointment, or dissatisfaction; and Neutral if it was primarily factual, interrogative, or lacked a dominant sentiment polarity.

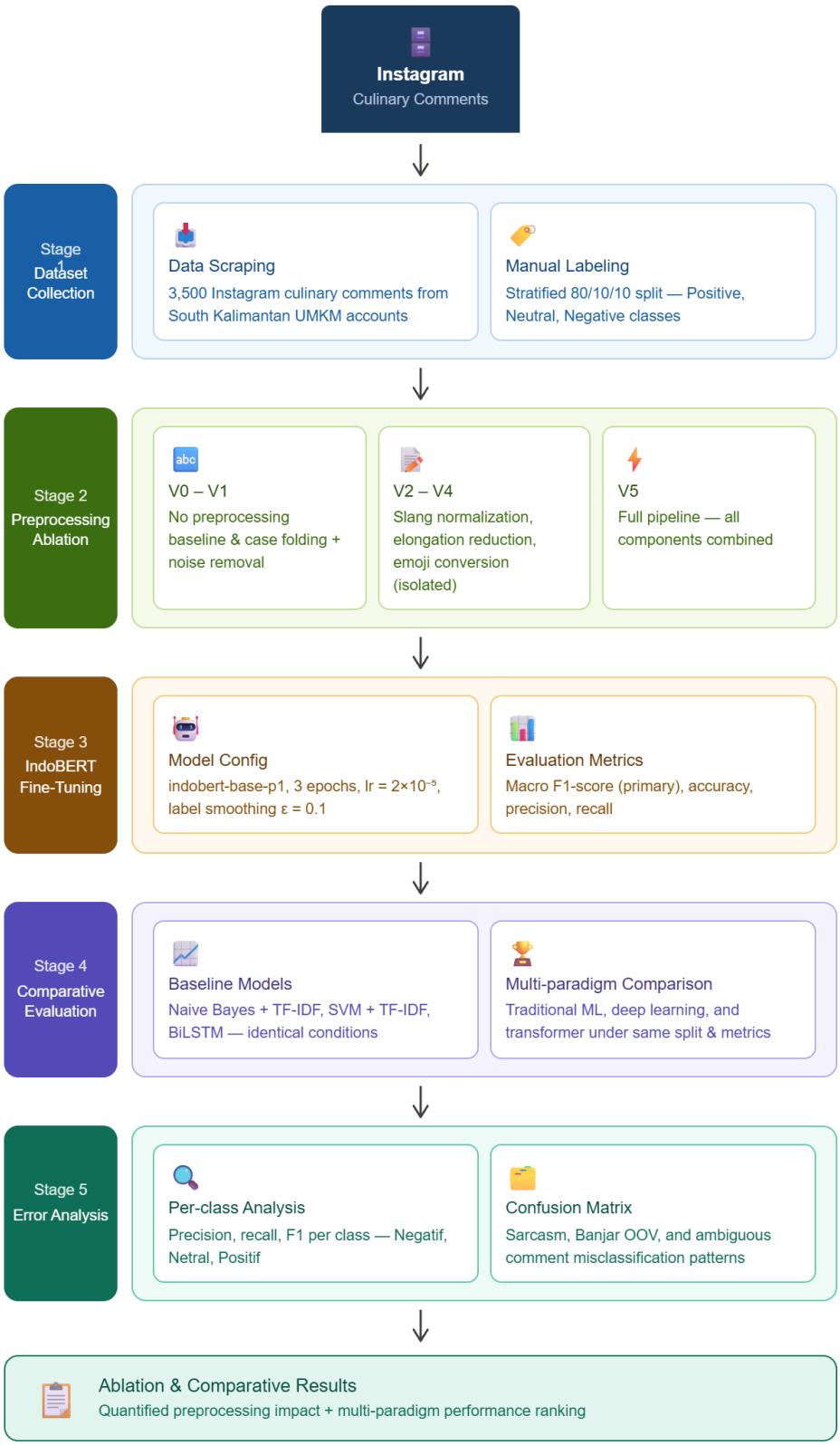

Figure 1. Overall experimental workflow of the study.

To assess label reliability, a validation stage was conducted on a randomly sampled 10% subset of the dataset (350 comments). Validation was performed by three independent

annotators with backgrounds in computer science and familiarity with Indonesian culinary discourse and the Banjar language. Each annotator independently assigned sentiment labels following the same annotation guideline. Disagreements were resolved through majority voting, where the label agreed upon by at least two annotators was adopted as the final label. Most sampled comments showed consistent agreement in the initial review, with disagreements primarily occurring in contextually ambiguous cases, such as sarcasm or mixed sentiment expressions.

It is acknowledged as a limitation that inter-annotator agreement was not formally quantified using a coefficient such as Cohen's Kappa, and that validation was performed on a subset rather than the full dataset. Future work should consider full-dataset multi-annotator validation with formal agreement metrics to further strengthen label reliability. Table 2 provides illustrative annotation examples.

Table 2. Annotation guideline with illustrative examples.

Label	Annotation guideline	Example comment
Positive	Compliment, satisfaction, or recommendation	"*Bagus kak*😍"
Neutral	Facts, questions, or no clear evaluation	"*Kak k Bekasi gmn*?"
Negative	Complaints, dissatisfaction, or rejection	"*Penyajian lama*"

The dataset was partitioned into training (80%), validation (10%), and test (10%) subsets using a stratified split with a random seed of 42, resulting in 2,800 training samples, 350 validation samples, and 350 test samples. Stratified sampling ensures that class proportions are preserved across all partitions. The test set was held out entirely during training and model selection and was used exclusively for final performance evaluation.

3.2. Preprocessing Ablation Design

3.2.1. Preprocessing Components

The preprocessing pipeline is designed to address the linguistic characteristics of informal Indonesian social media text, including non-standard spelling, slang, elongated characters, and emoji usage [5], [20]. The pipeline consists of four components applied sequentially, as described below:

P1 — Case folding and noise removal standardizes all text to lowercase and removes URLs, platform-specific symbols (e.g., mentions and hashtags while retaining their textual content), and punctuation-related noise through Unicode normalization. This serves as the baseline cleaning step applied to all variants.

P2 — Slang normalization replaces informal Indonesian terms, common abbreviations, and regional Banjar expressions with their standard Indonesian equivalents using a combined exact-match substitution dictionary constructed from two sources. The first source is a publicly available Indonesian slang corpus [35], from which 817 domain-relevant entries were selected, covering common abbreviations and informal variants such as "*gk*" → "*tidak*", "*bgt*" → "*banget*", and "*makasih*" → "*terima kasih*". The second source is the official Banjar–Indonesian dictionary published by the Indonesian Ministry of Education [36], from which 199 entries were selected, focusing on frequently observed culinary and daily-life expressions such as "*nyaman*" → "*enak*", "*padas*" → "*pedas*", "*masin*" → "*asin*", "*larang*" → "*mahal*", and "*masam*" → "*asam*". These sources were merged into a single dictionary of approximately 1,016 entries, with Banjar entries assigned higher lookup priority. Terms not found in the dictionary are left unchanged, resulting in potential out-of-vocabulary tokens, a limitation discussed in Section 4.5.

P3 — Elongation reduction normalizes repeated characters commonly used for emphasis, such as "enaaakkk" or "mantaaappp". The rule collapses characters repeated three or more times to at most two occurrences (e.g., "*enaaakkk*" → "*enakk*", "*mantaaappp*" → "*mantapp*"). Two occurrences are retained to preserve readability while reducing vocabulary fragmentation.

P4 − Emoji-to-token conversion converts selected sentiment-bearing emojis into their Indonesian lexical equivalents using a manually constructed dictionary (e.g., 😍 → "*sangat suka*", 👎 → "*jelek*", 😭 → "*sangat sedih*"). Emojis with ambiguous sentiment

polarity, such as 😂 and 🤣, are removed rather than converted, following the selective conversion strategy in [21].

Table 3. Preprocessing components with illustrative examples.

Step	Component	Raw input	Processed output
P1	Case folding + noise removal	"*ENAK BANGET*!!!"	"*enak banget*"
P2	Slang normalization	"*gk enak sih, porsinya dikit*"	"*tidak enak porsinya sedikit*"
P3	Elongation reduction	"*enaaakkk*!!!! *mantaaappp*"	"*enak mantap*"
P4	Emoji conversion	"😍😍"	"*sangat suka sangat suka*"

3.2.2. Ablation Variant Construction

Six preprocessing variants are constructed by systematically activating and deactivating individual components to isolate their contributions. Each variant is applied to the same dataset split, and each independently fine-tunes a freshly initialized IndoBERT-Base model under identical hyperparameter settings. Table 4 defines the composition and pipeline of each variant.

Table 4. Preprocessing variant composition and pipeline definition.

Variant	Description	P1	P2	P3	P4	Pipeline
V0_raw	No preprocessing	—	—	—	—	None
V1_base	Minimal cleaning only	✓	—	—	—	P1
V2_slang	Base + slang normalization	✓	✓	—	—	P1 + P2
V3_elonga	Base + elongation reduction	✓	—	✓	—	P1 + P3
V4_emoji	Base + emoji conversion	✓	—	—	✓	P1 + P4
V5_full	Full pipeline (all components)	✓	✓	✓	✓	P1 + P2 + P3 + P4

The ablation design follows a one-component-at-a-time (OFAT) approach, where V2 through V4 each activate exactly one additional component beyond the baseline V1_base, while V5 combines all components. This design enables the marginal contribution of each preprocessing step to be measured relative to V1_base [19], [22]. The V0_raw configuration serves as a no-preprocessing reference point against which all performance deltas are computed.

3.3. IndoBERT-Base Configuration

The primary model used in this study is IndoBERT-Base (indobenchmark/indobert-base-p1), a BERT-based transformer pre-trained on large-scale Indonesian corpora [13]. The architecture consists of 12 transformer encoder layers, 12 attention heads, a hidden dimension of 768, and approximately 124 million parameters. For three-class sentiment classification, a linear classification head is appended to the pooled [CLS] token representation from the final encoder layer. The logit vector for class prediction is computed as shown in Equation (1).

$$\hat{y} = W \cdot h_{[\mathrm{CLS}]} + b \tag{1}$$

where $h_{[\mathrm{CLS}]} \in \mathbb{R}^{768}$ is the pooled [CLS] representation from the final hidden layer, $W \in \mathbb{R}^{3\times 768}$ is the learned weight matrix of the classification head, $b \in \mathbb{R}^{3}$ is the bias vector, and $\hat{y} \in \mathbb{R}^{3}$ is the raw logit vector. The predicted class is $c^{*} = \mathrm{argmax}(\mathrm{softmax}(\hat{y}))$.

The model is optimized using cross-entropy loss with label smoothing ($\varepsilon = 0.1$) as defined in Equation (2).

$$\mathcal{L}_{\mathrm{LS}} = (1-\varepsilon)\cdot \mathcal{L}_{\mathrm{CE}} + \frac{\varepsilon}{\mathrm{K}} \sum_{k=1}^{\mathrm{K}} \mathcal{L}_{k} \tag{2}$$

where $\mathcal{L}_{\text{CE}}$ is the standard cross-entropy loss, $\varepsilon = 0.1$ is the smoothing factor, $K = 3$ is the number of classes, and $\mathcal{L}_k$ is the cross-entropy against the uniform distribution over all classes. Label smoothing reduces overconfidence in predictions during training and has been shown to improve generalization on short informal text [37]. The complete training hyperparameter configuration is provided in Table 5.

Table 5. Hyperparameters and training configuration..

Parameter	Value
Base model	indobenchmark/indobert-base-p1
Task	3-class classification (Positive/Neutral/Negative)
Max sequence length	128 tokens
Batch size	16
Gradient accumulation steps	1 (effective batch size = 16)
Learning rate	2×10^{-5}
Epochs	3
Warmup ratio	0.1
Weight decay	0.01
Dropout	0.1
Label smoothing (ε)	0.1
Early stopping patience	3 (threshold = 10^{-4})
Best checkpoint selection	Validation F1-weighted
Random seed	42

For each of the six ablation variants, the model is initialized from the same pre-trained weights and fine-tuned independently. The best checkpoint for each variant is selected based on validation F1-weighted performance and subsequently evaluated on the held-out test set. The choice of three training epochs follows the common recommendation that fine-tuning BERT-based models for 2–4 epochs is sufficient to achieve strong performance on downstream classification tasks [25]. This guideline is supported by subsequent studies [37], which show that model performance typically reaches a plateau after 2–3 epochs, with additional training increasing the risk of overfitting rather than improving generalization. This effect is more pronounced for short informal text [19].

An early stopping mechanism (patience = 3, threshold = 10^{-4}), monitored on validation F1-weighted) is employed to prevent overfitting if convergence occurs before the maximum number of epochs. Given the dataset size (2,800 training samples) and relatively short input sequences, three epochs represent a reasonable upper bound consistent with established practice in BERT fine-tuning for text classification [38].

3.4. Evaluation Metrics

Model performance is evaluated using four standard classification metrics: precision, recall, F1-score, and accuracy. Given the class imbalance in the dataset (Positive: 58.9%, Neutral: 26.6%, Negative: 14.5%), macro-averaged F1-score (F1-macro) is adopted as the primary evaluation metric, as it assigns equal weight to each class regardless of its support size and more accurately reflects performance on minority classes. Accuracy is reported as a supplementary metric for reference. These metrics are computed based on confusion matrix components: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) for each class. Accuracy, defined in Equation (3), measures the proportion of correctly classified instances:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

Precision, defined in Equation (4), measures the reliability of positive predictions:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

Recall, defined in Equation (5), measures the proportion of actual positive instances correctly identified:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

F1-score, defined in Equation (6), is the harmonic mean of precision and recall:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

Macro-averaged F1-score (F1-macro), defined in Equation (7), is used as the primary evaluation metric:

$$\text{F1-macro} = \frac{1}{K} \sum_{k=1}^{K} \text{F1}_k \tag{7}$$

where $K = 3$ and $F1_k$ denotes the F1-score for the k-th class. F1-macro assigns equal weight to each class regardless of its frequency, making it more informative than accuracy in imbalanced settings. Studies [31], [32] show that accuracy can remain deceptively high even when a model performs poorly on minority classes, whereas F1-macro better reflects such disparities. All metrics are computed on the held-out test set ($n = 350$), which is not used during training or model selection.

4. Results and Discussion

The results are presented in two stages. The multi-model comparison is reported first to establish the performance context within which the preprocessing ablation findings are interpreted. The ablation study, which constitutes the primary contribution of this work, is then presented to quantify the individual effect of each preprocessing component on IndoBERT fine-tuning performance.

4.1. Experimental Setup

All experiments were conducted in a Kaggle cloud computing environment equipped with an NVIDIA Tesla T4 GPU (16 GB VRAM). The software stack consisted of Python 3.12, PyTorch 2.10.0+cu128, Hugging Face Transformers 5.0.0, Datasets, Scikit-learn, and the Evaluate library. All random seeds were fixed at 42 across all frameworks to ensure reproducibility.

The IndoBERT-Base model was loaded from the indobenchmark/indobert-base-p1 checkpoint via the Hugging Face Hub. Each of the six preprocessing ablation variants was trained independently from the same pretrained checkpoint. Baseline models (Naive Bayes, SVM, and BiLSTM) were trained on the same fixed dataset split under an identical evaluation protocol.

4.2. Baseline Model Configurations and Multi-Model Comparison Results

To contextualize the IndoBERT-Base results, three baseline models were trained using the V5_full preprocessing output under the same dataset split and evaluation procedure. This ensures that performance differences across models reflect modeling capability rather than preprocessing variation. Naive Bayes with TF-IDF employs a Multinomial Naive Bayes classifier combined with TF-IDF feature extraction. The TF-IDF weight for a term in a document is defined as:

$$\text{tfidf}(t, d) = \text{tf}(t, d) \times \log\left(\frac{N}{\text{df}(t)}\right) \tag{8}$$

where $\text{tf}(t, d)$ is the raw term frequency of term t in document d, N is the total number of documents, and $\text{df}(t)$ is the number of documents containing t. The feature extractor uses unigram and bigram range (1,2) with a maximum vocabulary of 50,000 features. SVM with TF-IDF uses a LinearSVC classifier with the same TF-IDF feature space, where the regularization parameter is selected via grid search over {0.1, 1, 10} using validation F1-weighted as the selection criterion. BiLSTM is a bidirectional Long Short-Term Memory network with trainable word embeddings (embedding dimension 128, hidden size 128 per

direction, two stacked BiLSTM layers, dropout 0.3), trained using the Adam optimizer with a learning rate of 10^{-3} for a maximum of 10 epochs with early stopping (patience = 3, monitored: validation F1-weighted). Table 6 provides a consolidated summary of all model configurations.

Table 6. Baseline and primary model configurations.

Model	Feature type	Key settings	Selection criterion
Naive Bayes	TF-IDF (1,2)-gram	max_features = 50,000	Default
SVM	TF-IDF (1,2)-gram	LinearSVC, C ∈ {0.1, 1, 10}	Val F1-weighted
BiLSTM	Trainable embeddings	dim=128, hidden=128×2, dropout=0.3	Val F1-weighted
IndoBERT-Base	Contextual BERT	See Table 5	Val F1-weighted

Table 7 presents the comparative evaluation of all four models on the held-out test set using the V5_full preprocessing pipeline, along with overall performance metrics.

Table 7. Model comparison results on the test set (n = 350).

Model	F1-macro	F1-weighted	Accuracy	Precision	Recall
Naive Bayes + TF-IDF	0.6704	0.7331	0.7571	0.7677	0.7571
BiLSTM	0.7640	0.8113	0.8143	0.8113	0.8143
SVM + TF-IDF	0.7897	0.8288	0.8314	0.8306	0.8314
IndoBERT-Base	**0.8465**	**0.8793**	**0.8800**	**0.8790**	**0.8800**

The results in Table 7 establish a clear performance hierarchy across the three modeling paradigms, with IndoBERT-Base outperforming all baselines on every metric. Specifically, IndoBERT-Base surpasses SVM by +5.68 percentage points in macro F1-score (and +4.86 in accuracy), BiLSTM by +8.25 points in macro F1-score (and +6.57 in accuracy), and Naive Bayes by +17.61 points in macro F1-score (and +12.29 in accuracy).

The performance gap between IndoBERT-Base and SVM is particularly informative. Both models operate under the same V5_full preprocessing pipeline, and SVM has access to the same TF-IDF features, yet IndoBERT consistently achieves higher performance. This suggests that IndoBERT's advantage stems from its bidirectional contextual encoding, which enables the model to capture semantic relationships beyond surface-level token frequencies. This finding is consistent with [10], [37], which show that contextual representations learned by BERT-based models capture semantic patterns that cannot be reproduced by bag-of-words or static embedding approaches, regardless of preprocessing quality.

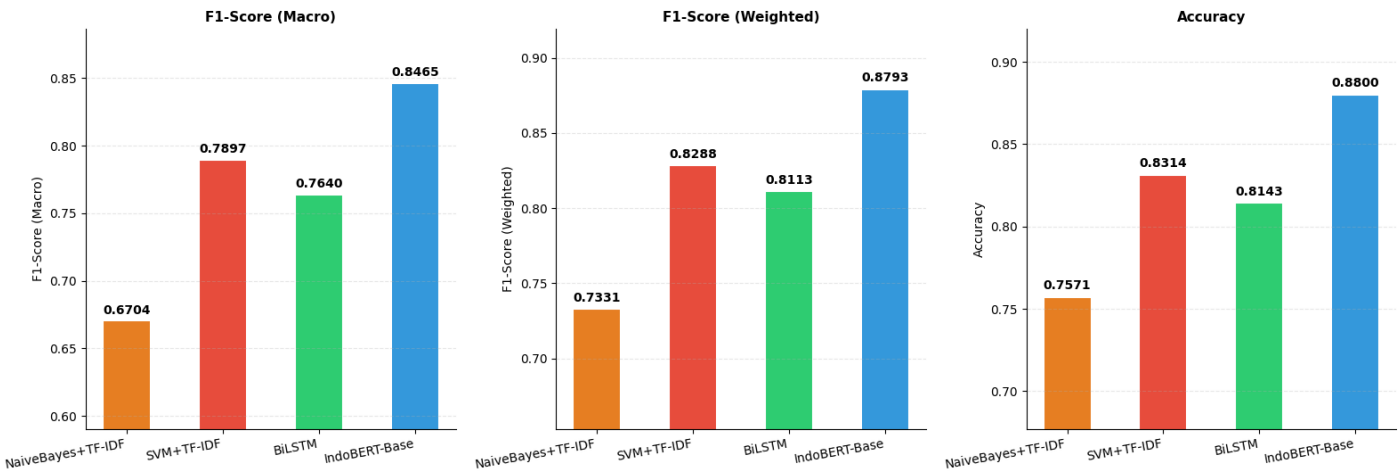

Figure 2. Performance comparison across model variants using the V5_full preprocessing pipeline under identical train/test splits, evaluated on the test set (n = 350).

The comparison between BiLSTM (F1-macro = 0.7640) and SVM (F1-macro = 0.7897) is also noteworthy, as SVM outperforms the neural baseline. This observation aligns with [30], which reports that for relatively small labeled datasets, traditional models with well-engineered features can match or exceed recurrent neural networks. With only 2,800 training samples, the BiLSTM model may lack sufficient data to effectively learn robust representations, whereas SVM benefits from a high-dimensional but fixed TF-IDF feature space.

It is important to interpret these results in the context of dataset size. The three modeling paradigms exhibit different sensitivities to training data volume: traditional classifiers (Naive Bayes and SVM) operate on fixed feature spaces and are relatively stable in low-data regimes, BiLSTM relies on learning representations from scratch and is more data-dependent, and IndoBERT-Base leverages pre-trained representations to mitigate limited labeled data. Consequently, the observed performance differences reflect the interaction between model architecture and data availability, and may not generalize to substantially larger datasets. These results should therefore be interpreted as indicative of relative performance under the specific conditions of this study rather than as a universal ranking of model effectiveness. Figure 2 provides a visual comparison of model performance across F1-macro, F1-weighted, and accuracy.

4.3. Preprocessing Ablation Study Results

Table 8 presents the complete ablation study results across all six preprocessing variants on the held-out test set, including delta scores computed relative to the V0_raw no-preprocessing baseline.

Table 8. Ablation study results on the test set (n = 350).

Variant	F1-macro	F1-weighted	Accuracy	Δ F1-macro	Δ Accuracy
V0_raw	0.7873	0.8248	0.8257	—	—
V1_base	0.8272	0.8627	0.8657	+0.0399	+0.0400
V2_slang	0.8482	0.8756	0.8771	+0.0609	+0.0514
V3_elonga	0.8283	0.8641	0.8657	+0.0410	+0.0400
V4_emoji	0.8235	0.8612	0.8629	+0.0362	+0.0371
V5_full	0.8465	0.8793	0.8800	+0.0592	+0.0543

The results reveal that preprocessing plays a non-uniform and structurally significant role in shaping IndoBERT fine-tuning performance, rather than acting as a marginal auxiliary step. First, the relatively strong baseline performance of V0_raw (F1-macro = 0.7873) indicates that IndoBERT's pre-trained contextual representations already encode substantial robustness to noisy informal input. This aligns with [19], which highlights the resilience of transformer-based models to input noise. However, the consistent improvements across all preprocessing variants suggest that this robustness is not sufficient for optimal performance, but instead provides a foundation that can be systematically enhanced. The performance gap of approximately +5.9 percentage points in F1-macro between V0_raw and V5_full demonstrates that preprocessing contributes not merely incremental gains, but practically meaningful improvements in downstream classification quality.

More importantly, the results show that preprocessing effectiveness is highly asymmetric across components. Slang normalization (V2_slang) yields the largest gain (+0.0609 F1-macro), substantially exceeding all other components. This indicates that the dominant source of performance degradation in informal Indonesian text is not noise in general, but lexical mismatch between user-generated vocabulary and the model's pre-training distribution. In practical terms, IndoBERT fails less due to noise itself, and more due to unrecognized lexical forms. By mapping slang and regional expressions into standardized tokens, preprocessing effectively "translates" the input into a space where the model's learned representations are applicable. This finding reframes preprocessing not as noise reduction, but as distribution alignment between training data and pre-training corpus.

Elongation reduction (V3_elonga) provides the second-largest improvement, but its impact is notably smaller than slang normalization. This suggests that character-level normalization addresses a secondary form of vocabulary fragmentation, where semantically valid tokens become unrecognizable due to surface variation. Unlike slang normalization, which resolves

semantic mismatches, elongation reduction primarily restores token identity, indicating that semantic normalization is more critical than orthographic normalization in this context.

Emoji conversion (V4_emoji) contributes the smallest gain, which provides an important insight: not all sentiment-bearing signals are equally exploitable by the model. While emojis encode affective information, their contribution appears limited when treated in isolation, it is consistent with [39]. This suggests that IndoBERT already captures sufficient sentiment cues from surrounding text, and that emoji signals act more as supplementary cues rather than primary carriers of sentiment. Additionally, the constrained emoji dictionary and the ambiguity of certain emojis further limit their effectiveness, reinforcing that preprocessing gains are bounded not only by method design but also by coverage and semantic clarity of the transformation.

The full pipeline (V5_full) achieves the best overall performance, but interestingly does not exceed the best single component (V2_slang) by a large margin. This reveals an important structural property: preprocessing gains are not strictly additive. Instead, the results exhibit a sub-additive pattern, suggesting partial redundancy between components. For instance, slang normalization may already resolve some elongated or informal variants, reducing the marginal contribution of subsequent steps. This indicates that preprocessing components interact in a non-linear manner, where the effectiveness of one step depends on the presence of others. Finally, the OFAT (one-component-at-a-time) design provides clear interpretability but inherently limits the exploration of interaction effects. As a result, the findings should be interpreted as identifying dominant individual contributors, rather than providing a complete optimization of preprocessing pipelines. Future work could extend this analysis toward interaction-aware designs to better capture combinatorial effects.

4.4. Per-Class Performance Analysis

Table 9 presents the per-class precision, recall, and F1-score for IndoBERT-Base with V5_full preprocessing on the test set, along with class support sizes reflecting the imbalanced distribution.

Table 9. Per-class classification report for IndoBERT-Base (V5_full), test set (n = 350).

Class	Support	Precision	Recall	F1-score
Negative	51	0.7755	0.7451	0.7600
Neutral	93	0.8778	0.8495	0.8634
Positive	206	0.9052	0.9272	0.9161
Macro avg	350	0.8528	0.8406	0.8465
Weighted avg	350	0.8790	0.8800	0.8793

The per-class results reveal that model performance is not only influenced by preprocessing design, but also strongly shaped by data distribution and decision bias during inference. The clear performance gradient—Positive > Neutral > Negative—mirrors the class imbalance in the dataset, but also reflects a deeper phenomenon: the model exhibits a systematic bias toward majority-class predictions.

The most critical observation lies in the Negative class. With an F1-score of 0.7600, it significantly underperforms compared to the Positive class (0.9161). More importantly, the lower recall (0.7451) relative to precision (0.7755) indicates that the model tends to underdetect negative sentiment rather than over-predict it. This suggests that when faced with ambiguous or weakly expressed sentiment, the model defaults toward safer majority classes (Positive or Neutral). This behavior is consistent with findings in [31] and [33], but also highlights a practical implication: errors are not symmetric across classes, and negative sentiment—often the most critical for business insights—is the most vulnerable.

This asymmetry also explains why accuracy alone is insufficient. Despite achieving an accuracy of 0.8800, the model still fails to adequately capture minority-class signals. The 15.61 percentage point gap between Positive and Negative F1-scores demonstrates that performance disparities are substantial and would remain hidden without macro-averaged evaluation. This reinforces the appropriateness of F1-macro as the primary metric, as recommended by [32], since it exposes performance imbalances that aggregate metrics obscure.

Beyond imbalance, these results also suggest that linguistic complexity is unevenly distributed across classes. Negative sentiment in informal Indonesian text is more likely to involve sarcasm, indirect expressions, or regional language usage, making it inherently more difficult to classify. In contrast, positive sentiment is often expressed more explicitly and consistently, which may explain its higher performance. This indicates that improving minority-class performance may require not only data balancing, but also targeted modeling of complex linguistic phenomena, such as sarcasm detection or regional language adaptation. Overall, the per-class analysis complements the ablation findings by showing that while preprocessing improves global performance, its benefits are not evenly distributed across classes, and model limitations remain most visible in minority and linguistically complex cases.

4.5. Error Analysis

Out of 350 test samples, IndoBERT-Base with V5_full preprocessing misclassified 42 instances, corresponding to an error rate of 12.0%. Figure 3 presents the confusion matrix with row-normalized percentages, illustrating the distribution of predictions for each true class. The diagonal cells (teal) represent correct predictions, with intensity proportional to frequency, while off-diagonal cells (coral) indicate misclassifications, with darker shades reflecting higher error concentration. The most prominent error pattern is the misclassification of Negatif as Positif (11 cases, 21.6% of true Negatif instances), which appears as the darkest off-diagonal cell. Table 10 provides representative qualitative examples of the dominant error categories.

The error analysis reveals three dominant and structurally meaningful patterns that account for most misclassifications, each pointing to distinct limitations of the current modeling and preprocessing pipeline. The most frequent error type is the misclassification of Negatif as Positif (11 cases, 26.2% of all errors). This pattern is largely driven by sarcasm and irony, which remain well-known challenges in sentiment analysis [23], [34]. In examples such as "tipisssss setipisss kulit bawang," the surface form lacks explicit negative lexical cues after preprocessing, leading the model to assign a positive label despite the underlying negative intent. As shown in [34], sarcastic expressions encode sentiment implicitly through pragmatics and context rather than explicit vocabulary, making them difficult to capture for models that rely primarily on lexical and contextual co-occurrence patterns. Even with fine-tuning, IndoBERT does not explicitly model pragmatic signals, which explains its susceptibility to such cases.

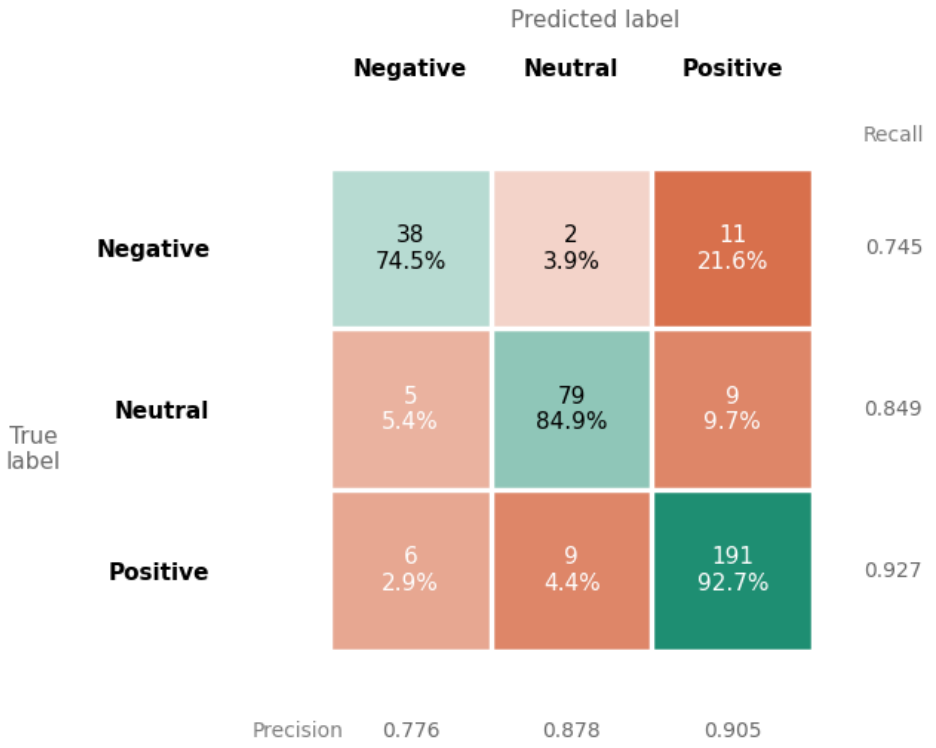

Figure 3. Confusion matrix of IndoBERT-Base with V5_full preprocessing (test set, n = 350).

Table 10. Qualitative error examples by misclassification category.

Category	Raw comment	Processed text	English interpretation	True label	Predicted	Confidence
Sarcasm	"@cibulkitchen *tipisssss setipisss kulit bawang wkwkwk*"	"cibulkitchen *tipiss setipiss kulit bawang wkwkwk*"	"[restaurant name] thin as onion skin hahaha" — a sarcastic complaint about portion size; "wkwkwk" indicates laughter (similar to "lol")	Negative	Positive	0.573
Banjar regional	"*parai tutup tarus*"	"*parai tutup tarus*"	"always closed" in Banjar; complaint about availability; "parai" (closed), "tarus" (always) are out-of-vocabulary	Negative	Neutral	0.900
Ambiguous short	"R*ed velvet dan tiramisu*"	"*red velvet dan tiramisu*"	Product listing without explicit sentiment; misclassified as positive due to association with menu items	Neutral	Positive	0.825
Code-mixing	"*Inggih jd ai tp ulun ni ke klinik dulu*"	"*inggih jd ai tp saya ni ke klinik dulu*"	"Yes, I will, but I need to go to the clinic first" — neutral statement with Banjar-Indonesian mix	Neutral	Positive	0.887
Positive with negative vocabulary	"*Kangennn kenapa kd buka di kamboja kaah*"	"*kangenn kenapa tidak buka di kamboja kaah*"	"I miss it, why don't you open in Kamboja?" — positive intent expressed with negation, misclassified as negative	Positive	Negative	0.740

The second major error source arises from regional language usage, particularly Banjar expressions that are not covered by the normalization dictionary or the pre-training corpus. For instance, the comment "parai tutup tarus" is misclassified as Netral with high confidence (0.900), as the key term "parai" is treated as an unknown token. This results in the loss of critical sentiment information at the representation level. Similar patterns are observed with other regional terms such as "lawas." This finding is consistent with [6], which shows that even multilingual and fine-tuned models exhibit performance degradation on low-resource regional languages. These errors highlight a structural limitation: preprocessing partially reduces vocabulary mismatch, but remains insufficient for handling unseen regional lexicons, which require either expanded dictionaries or dedicated multilingual adaptation.

The third pattern involves short and contextually ambiguous comments, such as "Red velvet dan tiramisu," which lack explicit sentiment signals but are nevertheless classified as Positif with high confidence (0.825). This behavior suggests that the model relies on distributional bias learned during training, where the presence of food-related terms is strongly associated with positive sentiment. Rather than reflecting semantic understanding, these predictions indicate a learned shortcut: culinary keywords are treated as proxies for positive sentiment. This type of error is therefore better interpreted as a form of dataset-induced bias rather than a limitation of the model architecture itself. Addressing this issue would require more balanced annotation or augmentation strategies that explicitly include neutral mentions of food items.

Across these categories, a common underlying theme emerges: misclassifications are not random but arise from systematic gaps between linguistic expression and model representation. Sarcasm introduces implicit sentiment that is not lexically encoded, regional language introduces tokens outside the model's vocabulary, and short ambiguous comments expose biases in learned associations. Together, these findings suggest that further improvements in performance will require not only better preprocessing, but also task-specific modeling strategies, such as sarcasm-aware training, regional language adaptation, or bias-aware data curation. Together, the error analysis complements the quantitative results by demonstrating that while preprocessing significantly improves global performance, residual errors are concentrated in linguistically complex and underrepresented cases. These patterns provide clear and actionable directions for future work, as further discussed in the following sections.

4.6. Synthesis of Key Findings

The experimental results across the three research objectives reveal a coherent pattern that extends beyond individual performance metrics and highlights the underlying mechanisms driving model behavior. The ablation study demonstrates that preprocessing components contribute unequally and through distinct mechanisms rather than acting as interchangeable steps. Slang normalization yields the largest improvement because it directly

addresses the mismatch between informal user-generated vocabulary and IndoBERT's pretraining distribution. In contrast, elongation reduction operates at the character level by restoring fragmented tokens, while emoji conversion introduces additional semantic cues with more limited impact. These differences indicate that preprocessing effectiveness is primarily determined by how directly a component aligns input representations with the model's learned vocabulary space, rather than by its ability to reduce noise in a general sense. Consequently, preprocessing should be understood as a form of distribution alignment, not merely text cleaning.

The model comparison results should be interpreted within this preprocessing context. The superior performance of IndoBERT-Base reflects not only its contextual encoding capability but also its ability to leverage normalized input representations more effectively than models operating on fixed feature spaces. At the same time, the observation that SVM outperforms BiLSTM highlights an important practical consideration: model effectiveness is conditioned by the interaction between architectural capacity and dataset size. With a relatively limited training set (2,800 samples), models that rely on learning representations from scratch may underperform compared to approaches that leverage stable feature spaces or pretrained knowledge. This suggests that model selection should be guided not only by architectural sophistication but also by data regime compatibility.

The per-class and error analysis further contextualize these findings by revealing that performance limitations are not uniformly distributed across classes. The lower performance on the Negatif class is influenced by both class imbalance and linguistic complexity. In particular, sarcasm and regional language usage represent two structurally distinct challenges: sarcasm introduces implicit sentiment that is not captured by lexical signals, while regional vocabulary introduces out-of-vocabulary tokens that remain unaddressed by the current preprocessing pipeline. These findings indicate that improvements in minority-class performance will likely require targeted modeling strategies, such as incorporating pragmatic cues or expanding lexical coverage, rather than further refinement of existing normalization steps.

Taken together, the results suggest that effective sentiment classification on informal Indonesian text requires a multi-layered approach, where preprocessing design, model architecture, and data characteristics are jointly considered. Rather than treating these components independently, the findings support a more integrated perspective in which preprocessing acts as a bridge between raw input variability and model representation capacity, while model selection and evaluation must account for dataset scale and class distribution.

5. Conclusions

This study investigated the impact of individual preprocessing components on IndoBERT-Base fine-tuning for Indonesian culinary sentiment classification and evaluated the resulting model against three baseline paradigms under controlled experimental conditions. The findings show that preprocessing is not a neutral or interchangeable step, but a critical factor that directly influences how effectively a pre-trained model can utilize its learned representations. Among the evaluated components, slang normalization provides the most substantial improvement, indicating that lexical mismatch between informal text and the pretraining corpus constitutes a primary performance bottleneck. Other components, including elongation reduction and emoji conversion, contribute additional gains through complementary mechanisms, resulting in the best performance when combined within a unified pipeline.

The comparative analysis further demonstrates that IndoBERT-Base achieves the strongest overall performance across all metrics, while also revealing that traditional models such as SVM can outperform neural architectures like BiLSTM under limited data conditions. This highlights the importance of aligning model choice with dataset scale and representation requirements rather than assuming consistent superiority of more complex architectures. From an evaluation perspective, the results confirm that macro-averaged F1-score provides a more informative assessment than accuracy in class-imbalanced settings, as it exposes disparities in class-level performance that aggregate metrics may obscure. The lower performance on the Negatif class, coupled with error patterns involving sarcasm, regional language usage, and ambiguous expressions, underscores the need for more targeted approaches to handling linguistically complex and underrepresented cases.

Several limitations should be acknowledged. The dataset size remains relatively modest, which constrains the generalizability of cross-paradigm comparisons and may favor models

that leverage pre-trained representations. In addition, the ablation design follows a one-component-at-a-time approach, which isolates individual effects but does not capture interaction effects or sensitivity to preprocessing order. These factors suggest that the reported results represent a structured but partial view of preprocessing effectiveness. Future work may extend this study by exploring interaction-aware ablation designs, expanding regional language coverage through region-specific resources or multilingual adaptation, incorporating imbalance-aware training strategies to improve minority-class recall, and integrating auxiliary tasks such as sarcasm detection. More broadly, extending the framework toward aspect-based sentiment analysis may provide more fine-grained insights into culinary user opinions. Overall, this study contributes empirical evidence that preprocessing plays a central and non-trivial role in transformer-based sentiment analysis for informal Indonesian text, and that its design should be treated as an integral component of the modeling process rather than a fixed preliminary step.

Author Contributions: Conceptualization: R.B. and S.W.S.; Methodology: R.B. and F.A.; Software: R.B.; Validation: S.W.S., and F.A.; Formal analysis: R.B. and A.F.; Investigation: R.B.; Resources: S.W.S.; Data curation: R.B. and F.A.; Writing original draft preparation: R.B.; Writing review and editing: R.B.; Visualization: R.B. and R.A.N.; Supervision: S.W.S. and A.F.; Project administration: S.W.S., F.A., R.A.N., and A.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The datasets in this study are publicly accessible on Kaggle at the following URL: https://www.kaggle.com/datasets/rahmatilkom/umkmkuliner.

Acknowledgments: The author gratefully acknowledges the Department of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Lambung Mangkurat, for the academic support provided during this research. Special appreciation is extended to the supervising lecturer for valuable guidance, constructive feedback, and continuous support throughout the research and manuscript preparation.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] S. Abdullah, P. Van Cauwenberge, H. Vander Bauwhede, and P. O'Connor, "Review Ratings, Sentiment in Review Comments, and Restaurant Profitability: Firm-Level Evidence," *Cornell Hosp. Q.*, vol. 65, no. 3, pp. 378–392, Aug. 2024, doi: 10.1177/19389655231214758.

[2] Y. Wang, J. Kim, and J. Kim, "The financial impact of online customer reviews in the restaurant industry: A moderating effect of brand equity," *Int. J. Hosp. Manag.*, vol. 95, p. 102895, May 2021, doi: 10.1016/j.ijhm.2021.102895.

[3] A. R. Putra, E. Ernawati, J. Jahroni, T. S. Anjanarko, and E. Retnowati, "Creative Economy Development Efforts in Culinary Business," *J. Soc. Sci. Stud.*, vol. 2, no. 1, pp. 21–26, Jan. 2022, doi: 10.56348/jos3.v2i1.17.

[4] H. Mulyono and A. R. Syamsuri, "Organizational Agility, Open Innovation, and Business Competitive Advantage: Evidence from Culinary SMEs in Indonesia," *Int. J. Soc. Sci. Bus.*, vol. 7, no. 2, pp. 268–275, Jun. 2023, doi: 10.23887/ijssb.v7i2.54083.

[5] A. F. Aji *et al.*, "One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 7226–7249. doi: 10.18653/v1/2022.acl-long.500.

[6] G. I. Winata *et al.*, "NusaX: Multilingual Parallel Sentiment Dataset for 10 Indonesian Local Languages," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2023, pp. 815–834. doi: 10.18653/v1/2023.eacl-main.57.

[7] M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," *Knowledge-Based Syst.*, vol. 226, p. 107134, Aug. 2021, doi: 10.1016/j.knosys.2021.107134.

[8] Y. Mao, Q. Liu, and Y. Zhang, "Sentiment analysis methods, applications, and challenges: A systematic literature review," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 36, no. 4, p. 102048, Apr. 2024, doi: 10.1016/j.jksuci.2024.102048.

[9] Y. Yanfi, Y. Heryadi, L. Lukas, W. Suparta, and Y. Arifin, "Sentiment Analysis of User Review on Indonesian Food and Beverage Group using Machine Learning Techniques," in *2022 IEEE Creative Communication and Innovative Technology (ICCIT)*, Nov. 2022, pp. 1–5. doi: 10.1109/ICCIT55355.2022.10118707.

[10] E. C. Garrido-Merchan, R. Gozalo-Brizuela, and S. Gonzalez-Carvajal, "Comparing BERT Against Traditional Machine Learning Models in Text Classification," *J. Comput. Cogn. Eng.*, vol. 2, no. 4, pp. 352–356, Apr. 2023, doi: 10.47852/bonviewJCCE3202838.

[11] R. Pramana, M. Jonathan, H. S. Yani, and R. Sutoyo, "A Comparison of BiLSTM, BERT, and Ensemble Method for Emotion Recognition on Indonesian Product Reviews," *Procedia Comput. Sci.*, vol. 245, pp. 399–408, 2024, doi: 10.1016/j.procs.2024.10.266.

[12] C.-H. Lin and U. Nuha, "Sentiment analysis of Indonesian datasets based on a hybrid deep-learning strategy," *J. Big Data*, vol. 10, no. 1, p. 88, May 2023, doi: 10.1186/s40537-023-00782-9.
[13] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 2020, pp. 843–857. doi: 10.18653/v1/2020.aacl-main.85.
[14] F. Koto, J. H. Lau, and T. Baldwin, "IndoBERTweet: A Pretrained Language Model for Indonesian Twitter with Effective Domain-Specific Vocabulary Initialization," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 10660–10668. doi: 10.18653/v1/2021.emnlp-main.833.
[15] H. Murfi, Syamsyuriani, T. Gowandi, G. Ardaneswari, and S. Nurrohmah, "BERT-based combination of convolutional and recurrent neural network for indonesian sentiment analysis," *Appl. Soft Comput.*, vol. 151, p. 111112, Jan. 2024, doi: 10.1016/j.asoc.2023.111112.
[16] H. Imaduddin, F. Y. A'la, and Y. S. Nugroho, "Sentiment Analysis in Indonesian Healthcare Applications using IndoBERT Approach," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 8, 2023, doi: 10.14569/IJACSA.2023.0140813.
[17] E. Yulianti and N. K. Nissa, "ABSA of Indonesian customer reviews using IndoBERT: single- sentence and sentence-pair classification approaches," *Bull. Electr. Eng. Informatics*, vol. 13, no. 5, pp. 3579–3589, Oct. 2024, doi: 10.11591/eei.v13i5.8032.
[18] R. I. Perwira, V. A. Permadi, D. I. Purnamasari, and R. P. Agusdin, "Domain-Specific Fine-Tuning of IndoBERT for Aspect-Based Sentiment Analysis in Indonesian Travel User-Generated Content," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 11, no. 1, pp. 30–40, Mar. 2025, doi: 10.20473/jisebi.11.1.30-40.
[19] M. Siino, I. Tinnirello, and M. La Cascia, "Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers," *Inf. Syst.*, vol. 121, p. 102342, Mar. 2024, doi: 10.1016/j.is.2023.102342.
[20] A. Bustamin, A. A. Prayogi, D. Siswanto, M. Rafrin, and A. Nurdin, "Text normalization for Indonesian slang words in sentiment analysis development," *ICIC Express Lett. Part B Appl.*, vol. 16, no. 2, pp. 121–129, Feb. 2025, doi: 10.24507/icicelb.16.02.121.
[21] A. Khan, D. Majumdar, and B. Mondal, "Sentiment analysis of emoji fused reviews using machine learning and Bert," *Sci. Rep.*, vol. 15, no. 1, p. 7538, Mar. 2025, doi: 10.1038/s41598-025-92286-0.
[22] M. Pota, M. Ventura, H. Fujita, and M. Esposito, "Multilingual evaluation of pre-processing for BERT-based sentiment analysis of tweets," *Expert Syst. Appl.*, vol. 181, p. 115119, Nov. 2021, doi: 10.1016/j.eswa.2021.115119.
[23] D. Suhartono, W. Wongso, and A. Tri Handoyo, "IdSarcasm: Benchmarking and Evaluating Language Models for Indonesian Sarcasm Detection," *IEEE Access*, vol. 12, pp. 87323–87332, 2024, doi: 10.1109/ACCESS.2024.3416955.
[24] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artif. Intell. Rev.*, vol. 55, no. 7, pp. 5731–5780, Oct. 2022, doi: 10.1007/s10462-022-10144-1.
[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North*, 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.
[26] H. Ahmadian, T. F. Abidin, H. Riza, and K. Muchtar, "Hybrid Models for Emotion Classification and Sentiment Analysis in Indonesian Language," *Appl. Comput. Intell. Soft Comput.*, vol. 2024, no. 1, Jan. 2024, doi: 10.1155/2024/2826773.
[27] D. R. I. M. Setiadi, W. Warto, A. R. Muslikh, K. Nugroho, and A. N. Safriandono, "Aspect-Based Sentiment Analysis on E-commerce Reviews using BiGRU and Bi-Directional Attention Flow," *J. Comput. Theor. Appl.*, vol. 2, no. 4, pp. 470–480, Apr. 2025, doi: 10.62411/jcta.12376.
[28] A. Bahmani, "Fusion of Statistical and Stylistic Text Features with SVM for Persian Sentiment Analysis," *J. Futur. Artif. Intell. Technol.*, vol. 2, no. 4, pp. 534–548, Dec. 2025, doi: 10.62411/faith.3048-3719-287.
[29] N. F. Adhim and N. Cahyono, "Optimization of IndoBERT for Sentiment Analysis of FOMO on Social Media Through Fine-Tuning and Hybrid Labeling," *J. Appl. Informatics Comput.*, vol. 9, no. 6, pp. 3786–3797, Dec. 2025, doi: 10.30871/jaic.v9i6.11686.
[30] A. Romadhony, S. Al Faraby, R. Rismala, U. N. Wisesty, and A. Arifianto, "Sentiment Analysis on a Large Indonesian Product Review Dataset," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 10, no. 1, pp. 167–178, Feb. 2024, doi: 10.20473/jisebi.10.1.167-178.
[31] S. Henning, W. Beluch, A. Fraser, and A. Friedrich, "A Survey of Methods for Addressing Class Imbalance in Deep-Learning Based Natural Language Processing," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2023, pp. 523–540. doi: 10.18653/v1/2023.eacl-main.38.
[32] J. Opitz, "A Closer Look at Classification Evaluation Metrics and a Critical Reflection of Common Evaluation Practice," *Trans. Assoc. Comput. Linguist.*, vol. 12, pp. 820–836, Jun. 2024, doi: 10.1162/tacl_a_00675.
[33] D. A. Kristiyanti, S. A. Sanjaya, V. C. Tjokro, and J. Suhali, "Dealing imbalance dataset problem in sentiment analysis of recession in Indonesia," *IAES Int. J. Artif. Intell.*, vol. 13, no. 2, p. 2060, Jun. 2024, doi: 10.11591/ijai.v13.i2.pp2060-2072.
[34] Y. Y. Tan, C.-O. Chow, J. Kanesan, J. H. Chuah, and Y. Lim, "Sentiment Analysis and Sarcasm Detection using Deep Multi-Task Learning," *Wirel. Pers. Commun.*, vol. 129, no. 3, pp. 2213–2237, Apr. 2023, doi: 10.1007/s11277-023-10235-4.
[35] N. Aliyah Salsabila, Y. Ardhito Winatmoko, A. Akbar Septiandri, and A. Jamal, "Colloquial Indonesian Lexicon," in *2018 International Conference on Asian Language Processing (IALP)*, Nov. 2018, pp. 226–229. doi: 10.1109/IALP.2018.8629151.
[36] Y. Puspita Sari, A. Husna, E. Anggraini, and S. Akbari, "Kamus: bahasa Banjar-Indonesia untuk pelajar," *Kementerian Pendidikan Dasar dan Menengah Republik Indonesia*. https://repositori.kemendikdasmen.go.id/35353/
[37] N. J. Prottasha *et al.*, "Transfer Learning for Sentiment Analysis Using BERT Based Supervised Fine-Tuning," *Sensors*, vol. 22, no. 11, p. 4157, May 2022, doi: 10.3390/s22114157.
[38] N. D. A. Saputra, M. Muljono, A. Karim, and D. R. I. M. Setiadi, "End-to-End Fine-Tuning of DeBERTa-Base for Stance Detection," *J. Futur. Artif. Intell. Technol.*, vol. 2, no. 4, pp. 698–715, Feb. 2026, doi: 10.62411/faith.3048-3719-168.
[39] J. Chen, Z. Yao, S. Zhao, and Y. Zhang, "Fusion Pre-trained Emoji Feature Enhancement for Sentiment Analysis," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 22, no. 4, pp. 1–14, Apr. 2023, doi: 10.1145/3578582.