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

Research Article

Language-Similarity-Guided Transfer Fine-Tuning of Pre-trained Transformer Models for Sentiment Analysis Across 12 Indonesian Regional Languages

Brian Rizqi Paradisiaca Darnoto * and Dony Bahtera Firmawan

Informatics Department, University of Jember, Jember 68121, Indonesia;
e-mail : brianrizqi@unej.ac.id; donybf@unej.ac.id
* Corresponding Author : Brian Rizqi Paradisiaca Darnoto

Abstract: Sentiment analysis for Indonesian regional languages faces two persistent challenges: labeled training data is extremely limited for most regional varieties, and transformer models pre-trained on Bahasa Indonesia do not generalize reliably to languages with substantially different morphological structures. Prior work on the NusaX benchmark has primarily relied on direct fine-tuning, treating each regional language independently and without exploiting linguistic proximity between related languages as a transfer signal. This paper proposes Language-Similarity-Guided Transfer (LSGT), a sequential fine-tuning strategy that first adapts a pre-trained model to a pivot language selected using character trigram similarity, followed by fine-tuning on the target language. Four transformer models are evaluated across all 12 NusaX languages using the official train/validation/test splits: IndoBERT, NusaBERT, mBERT, and XLM-R. Performance is evaluated using four metrics: accuracy, macro F1, macro precision, and macro recall. Experimental results show that LSGT improves macro F1 in 44 of 48 model-language combinations, demonstrating that the fine-tuning strategy itself is a major factor in low-resource cross-lingual sentiment classification. XLM-R benefits most strongly from LSGT, achieving an average improvement of +0.137 macro F1 and a peak gain of +0.298 on Madurese. SHAP-based token attribution analysis further reveals that predictions rely heavily on named entities and domain-specific nouns rather than sentiment-bearing vocabulary, indicating a dataset-level bias inherited from the original SmSA corpus and propagated through the NusaX translation pipeline.

Keywords: Indonesian regional languages; Low-resource NLP; NusaX; Pre-trained language models; Sentiment analysis; SHAP explainability; Transfer learning; XLM-R.

Received: April, 9th 2026
Revised: May, 5th 2026
Accepted: May, 5th 2026
Published: May, 7th 2026

1. Introduction

Sentiment analysis classifies text into positive, negative, or neutral polarity and remains one of the most extensively studied tasks in natural language processing (NLP) [1]. Its applications span social media monitoring, product review analysis, political discourse tracking, and public opinion mining [2]. Despite substantial progress in high-resource languages, the field remains geographically imbalanced: most datasets, benchmarks, and pre-trained models are concentrated in English, Chinese, and a small number of other well-resourced languages, while languages spoken by smaller or underrepresented communities receive far less attention [3].

Indonesia illustrates this imbalance clearly. With more than 700 regional languages, it represents one of the most linguistically diverse countries in the world [4], [5]. Nevertheless, NLP research in Indonesia has focused almost entirely on Bahasa Indonesia, the national language. Regional languages such as Javanese, Sundanese, Buginese, and Toba Batak remain largely absent from the computational linguistics literature despite their substantial speaker populations. Javanese alone, with approximately 82 million speakers, exceeds the population of many European languages that already possess mature NLP ecosystems. For regional governments, healthcare providers, and local media organizations operating in areas where

Bahasa Indonesia is not the dominant daily language, sentiment analysis systems that cannot process regional languages have limited practical utility.

The central technical challenge is low-resource adaptation. Transformer-based fine-tuning has become the dominant paradigm for sentiment classification, but its effectiveness depends heavily on the availability of labeled data in the target language. For most Indonesian regional languages, such resources are extremely limited. The NusaX benchmark [6], currently the most comprehensive multilingual Indonesian sentiment dataset, provides roughly 1,000 labeled samples per language, with only 700 available for training under the official split. This scale is substantially below what standard fine-tuning workflows typically assume, causing models to adapt from poorly aligned initial representations. Consequently, the effectiveness of the fine-tuning strategy itself becomes as important as the underlying model architecture. One promising direction is to exploit linguistic similarity between related regional languages, using structurally similar languages as intermediate transfer steps rather than treating each target language independently.

This study evaluates four pre-trained transformer models representing distinct design choices for Indonesian and multilingual NLP. BERT [7] introduced masked bidirectional pre-training for downstream NLP tasks. IndoBERT [8], trained on approximately 4 billion Indonesian words, established strong baselines across Indonesian NLP benchmarks. NusaBERT [9] extends IndoBERT through continued pre-training on corpora covering 13 Indonesian regional language varieties. In the multilingual setting, mBERT [10] and XLM-R [11] both support more than 100 languages, although XLM-R was trained on a substantially larger CommonCrawl corpus, resulting in stronger cross-lingual representations for many low-resource scenarios. Importantly, XLM-R has not previously been evaluated on the NusaX sentiment benchmark. Taufiq and Sutopo [12] conducted the most comprehensive prior evaluation on NusaX, comparing eight BERT-based models across 10 regional languages and showing that Indonesian-specific models consistently outperform multilingual alternatives, while Toba Batak remains the most difficult language across all configurations.

Two methodological gaps remain insufficiently explored. The first concerns transfer strategy. Existing studies on NusaX rely on direct fine-tuning, where a pre-trained checkpoint is adapted independently to each target language. This assumption ignores the substantial lexical and morphological overlap between many Indonesian regional languages. Intermediate-task transfer learning has repeatedly been shown to improve downstream cross-lingual performance [13], [14], particularly when the intermediate language is closely related to the target language. The distinction introduced in this work is that the intermediate pivot language is selected systematically using character trigram similarity rather than manual intuition or arbitrary selection. To the best of our knowledge, no prior work has applied a reproducible similarity-guided pivot selection strategy to sequential fine-tuning on the NusaX benchmark.

The second gap concerns interpretability. Performance metrics alone provide limited insight into what models actually learn. A macro F1 score on a regional language does not indicate whether the model relies on sentiment-bearing vocabulary or merely memorizes named entities correlated with sentiment labels in the training corpus. This distinction is important for understanding model generalization and diagnosing language-specific performance limitations. While prior work on NusaX [12] reports classification performance, it does not analyze the linguistic features driving model predictions. SHAP [15] enables model-agnostic token-level attribution analysis, allowing inspection of whether predictions are driven primarily by sentiment expressions, named entities, or domain-specific vocabulary. Such analysis has not previously been conducted on the NusaX sentiment benchmark.

This paper addresses both gaps through Language-Similarity-Guided Transfer (LSGT), a sequential fine-tuning strategy in which a pivot language is selected using character trigram similarity before adaptation to the target language. The study evaluates IndoBERT, NusaBERT, mBERT, and XLM-R across all 12 NusaX languages using the official dataset splits and four evaluation metrics: accuracy, macro F1, macro precision, and macro recall. In addition to extending prior work through the inclusion of XLM-R and a broader evaluation framework, the study applies SHAP attribution analysis to examine how transfer learning affects token-level decision behavior, particularly in consistently difficult languages. The remainder of this paper is organized as follows: Section 2 reviews related work, Section 3 describes the methodology, Section 4 presents the experimental results and discussion, and Section 5 concludes the paper.

2. Related Work

2.1. Sentiment Analysis in Indonesian and Low-Resource Settings

Research on Indonesian sentiment analysis has progressed through three major methodological phases, each addressing limitations of the previous generation of approaches. Early studies relied on classical machine learning methods such as Naive Bayes, support vector machines, and random forests [16], [17], primarily evaluated on Twitter and product-review datasets. These approaches depended heavily on manually constructed sentiment lexicons, which were limited even for Bahasa Indonesia and effectively nonexistent for regional languages. As a result, out-of-vocabulary expressions, colloquial terms, negation, and code-switching frequently degraded classification performance.

Around 2019, recurrent neural architectures began replacing lexicon-based approaches by learning contextual representations directly from data. LSTM [18] and GRU models improved the handling of contextual dependencies relevant to sentiment prediction, particularly negation and irony in longer texts. Purwarianti and Crisdayanti [19] demonstrated substantial gains over classical baselines on the SmSA corpus, which later became the source dataset for NusaX. Nevertheless, recurrent architectures introduced new limitations: performance on short or code-switched texts remained unstable [20], and these models were never evaluated systematically on Indonesian regional languages.

Transformer-based models substantially improved contextual modeling by encoding full-sequence dependencies through self-attention mechanisms. Murfi et al. [21] combined BERT with convolutional and recurrent layers and showed that transformer-based representations generalized better across domains than recurrent-only architectures. Ahmadian et al. [22] further demonstrated that hybrid IndoBERT architectures outperform standard fine-tuning on IndoNLU benchmark tasks. Fine-tuned IndoBERT consistently surpassed classical baselines in applications such as mobile app reviews [23], law-enforcement sentiment analysis on Twitter [24], and electoral discourse analysis [25], often exceeding 83% classification accuracy. However, a common limitation across these studies is that all experiments were conducted exclusively in Bahasa Indonesia. Regional language sentiment analysis, particularly under severe data scarcity, remains largely unexplored.

Research specifically targeting Indonesian regional languages is still limited. Aji et al. [5] reported that fewer than a dozen of the more than 700 Indonesian languages possess any labeled NLP resources. NusaX [6], introduced in 2023, represents the first large-scale multilingual benchmark for Indonesian regional language NLP, created by professionally translating SmSA examples into 10 regional languages. Taufiq and Sutopo [12] evaluated eight BERT-based models on these languages using direct fine-tuning and found that Indonesian-specific models consistently outperform multilingual models, while Toba Batak remains the most difficult language across all configurations. Their study, however, used a custom 80/20 split and reported only accuracy, limiting direct comparability with the official NusaX benchmark. Furthermore, no prior work has evaluated XLM-R on NusaX, explored similarity-guided intermediate fine-tuning across regional languages, or applied token-level attribution analysis to explain language-specific performance differences. These gaps motivate the present study.

2.2. Pre-trained Transformer Models for Indonesian

BERT [7] established the foundation of modern transfer learning in NLP by introducing bidirectional Transformer pre-training using masked language modeling and next sentence prediction. The Transformer architecture [26] underlying BERT employs multi-head self-attention to model long-range contextual dependencies more effectively than recurrent architectures.

IndoBERT [8] adapted this paradigm to Indonesian by pre-training on the Indo4B corpus, a 23 GB collection of Indonesian news articles, Wikipedia pages, subtitles, and social media text. The model established state-of-the-art performance across multiple IndoNLU benchmark tasks. A separate IndoBERT checkpoint developed within the IndoLEM project [27] was trained on a 220-million-word Indonesian corpus and has also been widely used in downstream evaluations.

NusaBERT [9] extends IndoBERT through continued pre-training on multilingual corpora covering 13 Indonesian regional language varieties, including data from Wikipedia, CulturaX [28][12

, and the NLLB dataset. The model was specifically designed to improve representation quality for regional languages. Nevertheless, Taufiq and Sutopo [12] reported that NusaBERT-base underperformed IndoBERT-base across all evaluated NusaX languages. One plausible explanation is that the regional-language corpora used during continued pre-training were insufficient to produce strongly discriminative representations, while multilingual mixing may have weakened calibration for standard Indonesian representations. This study evaluates the updated LazarusNLP/NusaBERT-base checkpoint, which incorporates revised training data.

Multilingual BERT (mBERT) [10] was pre-trained on Wikipedia corpora covering 104 languages using a shared WordPiece vocabulary. Pires et al. [10] showed that mBERT learns transferable cross-lingual representations even without explicit multilingual alignment objectives. XLM-R [11], trained on a CommonCrawl corpus approximately 15 times larger than mBERT's training data, demonstrated substantially stronger performance on cross-lingual benchmarks, particularly for low-resource languages. Kumar and Albuquerque [29] further confirmed the effectiveness of XLM-R for sentiment classification in low-resource Indian languages, while broader comparative studies identified XLM-R as one of the strongest multilingual baselines for cross-lingual sentiment analysis [30]. Despite these findings, XLM-R has not previously been evaluated on the NusaX sentiment benchmark.

2.3. Transfer Learning and Intermediate Fine-Tuning

Transfer learning from pre-trained Transformer models is now the dominant paradigm in NLP, but downstream performance depends heavily on how fine-tuning is structured. Direct fine-tuning adapts a pre-trained checkpoint directly to the target task using only target-language data. While straightforward, this approach becomes suboptimal in low-resource settings because the model initialization may remain poorly aligned with the target distribution. Peters et al. [31] showed that full fine-tuning generally outperforms feature extraction, although its effectiveness depends strongly on the similarity between the pre-training and downstream distributions. When labeled data is scarce, direct fine-tuning often fails to exploit the full potential of pre-trained representations.

Intermediate-task fine-tuning provides an alternative transfer strategy. Phang et al. [13] demonstrated that fine-tuning on a related intermediate task before the target task consistently improves cross-lingual transfer performance by reshaping representations toward more transferable semantic structures. Similar findings have been reported in cross-lingual transfer learning across languages. Wu and Dredze [14] showed that transfer from linguistically related source languages consistently outperforms transfer from distant languages, while Ruder et al. [32] identified lexical and morphological similarity as reliable predictors of transfer effectiveness. Barnes et al. [33] further confirmed this phenomenon in sentiment analysis, showing that bilingual sentiment representations outperform monolingual models under low-resource conditions.

Despite these findings, prior work provides limited guidance on how intermediate pivot languages should be selected. Existing studies typically rely on manually chosen source languages or fixed source-target transfer settings. The proposed Language-Similarity-Guided Transfer (LSGT) framework addresses this limitation by selecting the pivot language algorithmically using character trigram Jaccard similarity across all candidate languages. This makes the transfer process systematic, reproducible, and adaptable to different language families. A related finding in stance detection further emphasizes the importance of fine-tuning strategy itself: Saputra et al. [34] achieved 96.28% accuracy on a three-class stance classification task primarily through fine-tuning optimization rather than architectural modification. These results reinforce the argument that transfer strategy can be as important as model architecture in low-resource classification tasks. Table 1 summarizes the differences between direct fine-tuning, arbitrary intermediate fine-tuning, and the proposed LSGT framework.

Within the Indonesian regional-language context, the NusaX study [6] already showed that zero-shot transfer performs better between languages with higher vocabulary overlap, consistent with broader transfer-learning literature. However, no prior work has implemented a systematic pivot-selection strategy with sequential fine-tuning on this benchmark. Parameter-efficient fine-tuning methods, including adapter modules and low-rank adaptation approaches [35], are also relevant for low-resource NLP, although they address a different dimension of the problem. LSGT focuses on improving transfer through the ordering and

selection of fine-tuning stages rather than reducing trainable parameters, and both approaches could potentially be combined in future work.

Table 1. Comparison of fine-tuning paradigms for low-resource language sentiment analysis.

Approach	Pivot selection	Uses inter-language transfer?	Key limitation
Direct fine-tuning	None	No	Poor initialization for low-resource targets
Intermediate fine-tuning (arbitrary pivot)	Human intuition or fixed source	Partially	Pivot selection is unsystematic and difficult to generalize
LSGT (this study)	Character trigram Jaccard similarity (automatic)	Yes — systematic ity is measured only at the character level	Depends on pivot quality; similar-

2.4. Explainability in NLP

As transformer-based models have increasingly been deployed in high-impact settings, several families of explanation methods have emerged, each with different trade-offs in fidelity, interpretability, and computational cost. Gradient-based attribution methods [36] estimate token importance by computing the gradient of the model output with respect to the input representation. Although computationally efficient, these methods can assign high importance to tokens whose actual contribution to the prediction is unstable or context-dependent. Attention-based explanations instead interpret attention weights as indicators of token importance, providing interpretability without additional computation. However, Jain and Wallace [37] demonstrated that attention weights are not always faithful explanations, since high attention on a token does not necessarily imply strong causal influence on the model prediction. While Wiegreffe and Pinter [38] partially challenged this conclusion, the broader explainability literature [39] generally supports the use of model-agnostic perturbation-based approaches for more reliable post-hoc interpretation.

SHAP [15] approaches explainability from a cooperative game-theoretic perspective by assigning each feature an additive contribution score derived from Shapley values. In text classification, SHAP estimates how much each token changes the predicted probability when masked under a perturbation-based framework. Lundberg and Lee [15] showed that SHAP provides more faithful and theoretically grounded explanations than gradient- or attention-based methods across several benchmarks. Prior work has applied SHAP-style attribution to Indonesian sentiment analysis at the document level [21], while related explainability frameworks combining linguistic and behavioral signals have demonstrated improved interpretability in social-media depression detection [40]. However, no prior study has systematically applied token-level SHAP attribution to analyze model behavior across Indonesian regional languages in the NusaX benchmark.

SHAP is selected in this study because its perturbation-based masking strategy is model-agnostic, does not require gradient access, and provides token-level attribution scores that are generally more faithful than raw attention weights for Transformer-based models [39].

2.5. The NusaX Dataset

The NusaX [6] is a multilingual parallel sentiment corpus covering 12 languages: Indonesian, English, Acehnese, Balinese, Banjarese, Buginese, Javanese, Madurese, Minangkabau, Ngaju, Sundanese, and Toba Batak. The dataset was constructed by professionally translating selected examples from the SmSA sentiment corpus into 10 Indonesian regional languages. Each language contains the same label distribution: 383 negative, 239 neutral, and 378 positive samples. Because the source corpus originates from Indonesian restaurant and mobile application reviews, many named entities and domain-specific references are preserved across translations. Consequently, models may learn correlations between specific entities and sentiment labels that transfer consistently across language versions of the same example. This characteristic is particularly relevant for interpretability analysis, since token-level attribution methods may reveal whether predictions are driven by sentiment-bearing vocabulary or by translated named entities inherited from the source corpus.

Winata et al. [6] further reported substantial vocabulary overlap between several Indonesian regional languages, particularly between Javanese and Sundanese, while Indonesian itself shares relatively high character-level similarity with many regional varieties. These observations motivate the similarity-guided transfer strategy explored in this study and provide the empirical basis for the pivot-selection mechanism used in LSGT. Details of the experimental split configuration and its use in the proposed framework are provided in Section 3.1.

3. Methodology

Figure 1 presents the overall LSGT framework and experimental pipeline as a four-phase workflow. Phase 1 prepares the NusaX dataset using the official train/validation/test splits for all 12 languages (Section 3.1). Phase 2 computes pairwise character trigram similarity scores to construct the similarity matrix used for pivot-language selection (Section 3.3). Phase 3 performs model adaptation under two parallel configurations: standard direct fine-tuning (Section 3.4) and the proposed Language-Similarity-Guided Transfer (LSGT) strategy (Section 3.5). Finally, Phase 4 evaluates all model-language combinations using four classification metrics and applies SHAP attribution analysis to examine token-level decision behavior (Sections 3.6–3.7). The transformer models used in both fine-tuning settings are described in Section 3.2.

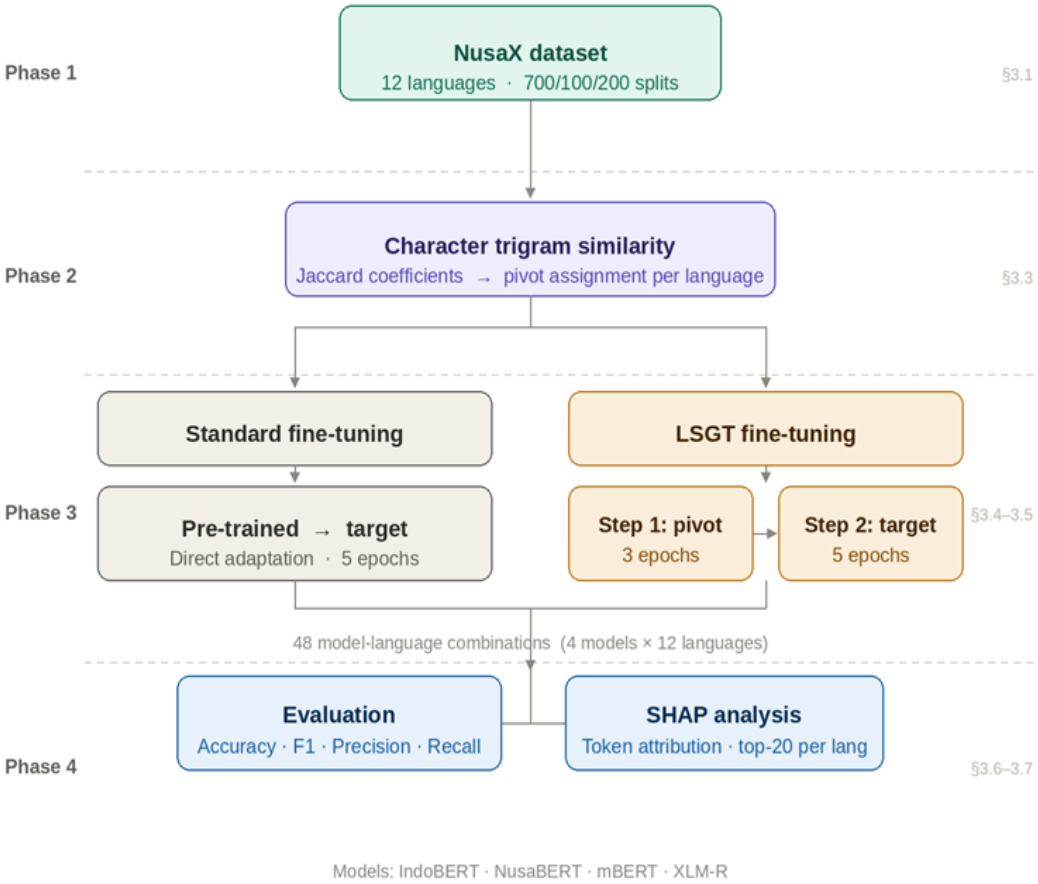

Figure 1. LSGT Framework and Experimental Pipeline

3.1 Dataset

This study uses the official NusaX sentiment dataset split consisting of 700 training samples, 100 validation samples, and 200 test samples per language. Using the official split ensures that the reported results are directly reproducible and comparable with the public NusaX benchmark. This configuration differs from the custom 80/20 partition used by Taufiq and Sutopo [12], which allocated approximately 800 training samples per language. Table 2 lists the 12 languages included in NusaX together with their ISO language codes. These languages form the candidate pool for the pivot-selection process used in the proposed LSGT framework.

Table 2. Languages included in the NusaX benchmark.

Language	Code	Language	Code
Indonesian	ind	Javanese	jav
English	eng	Madurese	mad
Acehnese	ace	Minangkabau	min
Balinese	ban	Ngaju	nij
Banjarese	bjn	Sundanese	sun
Buginese	bug	Toba Batak	bbc

3.2. Pre-trained Models

Four pre-trained Transformer models are evaluated in this study: IndoBERT, NusaBERT, mBERT, and XLM-R. These models represent different pre-training strategies for Indonesian and multilingual NLP, ranging from monolingual Indonesian pre-training to large-scale multilingual representation learning. IndoBERT uses the indobenchmark/indobert-base-p1 checkpoint with 124M parameters, pre-trained on approximately 4 billion Indonesian words. NusaBERT uses the LazarusNLP/NusaBERT-base checkpoint with 111M parameters, continuing IndoBERT pre-training on multilingual corpora covering 13 Indonesian regional language varieties. mBERT uses the bert-base-multilingual-cased checkpoint with 167M parameters, pre-trained on Wikipedia corpora spanning 104 languages. XLM-R uses the xlm-roberta-base checkpoint with 278M parameters, pre-trained on approximately 2.5 TB of CommonCrawl data covering 100 languages. Table 3 summarizes the evaluated models, including checkpoint identifiers, parameter counts, and pre-training corpora.

Table 3. Pre-trained transformer models evaluated in this study

Model	HuggingFace ID	Parameters	Pre-training data
IndoBERT	indobenchmark/indobert-base-p1	124M	4B Indonesian words
NusaBERT	LazarusNLP/NusaBERT-base	111M	IndoBERT + 13 regional languages (16B tokens)
mBERT	bert-base-multilingual-cased	167M	Wikipedia, 104 languages
XLM-R	xlm-roberta-base	278M	CommonCrawl, 100 languages (2.5 TB)

3.3. Linguistic Similarity Matrix

Pivot-language selection in LSGT is based on character trigram similarity between language pairs. Similarity is computed using the Jaccard coefficient over sets of character trigrams extracted from 500-word samples drawn from the training split of each language. Character-level n-gram similarity is suitable for Indonesian regional languages because it captures shared orthographic and morphological patterns common in Austronesian language families.

Character trigrams are preferred over phonological similarity measures because phonological resources are unavailable for most NusaX languages. They are also preferred over token-level lexical overlap, which is sensitive to tokenization schemes and vocabulary-size variation across different Transformer models. The 500-word sampling threshold was selected as a balance between representativeness and computational efficiency. Preliminary analysis showed that pairwise Jaccard coefficients stabilized within approximately ± 0.01 beyond 400 words, indicating that 500-word samples provide sufficiently stable similarity estimates.

Figure 2 presents the resulting similarity matrix. Each row corresponds to a target language, while the highest off-diagonal similarity score determines the pivot language assigned to that target. Indonesian serves as the pivot for most regional languages because it exhibits the highest average similarity across the language set and is commonly shared within multilingual speaker communities. Javanese is selected as the pivot for Balinese and Sundanese, consistent with their higher lexical and morphological overlap. Formally, pivot assignment is completed before model training begins and depends exclusively on linguistic similarity rather than downstream model performance.

3.4. Standard Fine-Tuning Baseline

The baseline configuration follows a standard direct fine-tuning strategy similar to that used by Taufiq and Sutopo [12]. Each pre-trained model is adapted directly to the target language without any intermediate transfer stage. A three-class classification head is added to the encoder output, and all parameters are fine-tuned end-to-end. Table 4 summarizes the hyperparameter configuration used across both the baseline and LSGT settings.

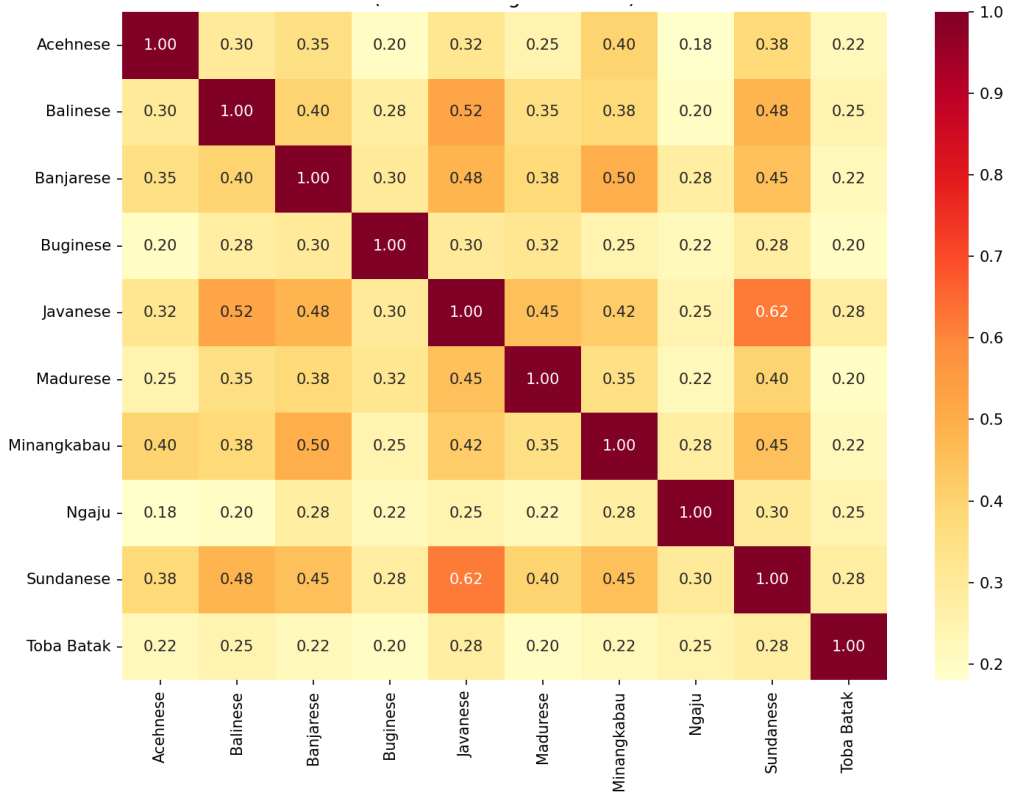

Figure 2. Linguistic similarity matrix based on character trigram Jaccard coefficients

Table 4. Hyperparameter configuration for fine-tuning

Hyperparameter	Value
Epochs (target language)	5
Epochs (pivot language, LSGT only)	3
Batch size	32
Learning rate	2e-5
Optimizer	AdamW
Weight decay	0.01
Learning-rate scheduler	Cosine with warmup (ratio 0.1)
Maximum sequence length	128 tokens
Mixed precision	FP16/FP32

3.5. Language-Similarity-Guided Transfer (LSGT)

Language-Similarity-Guided Transfer (LSGT) introduces an intermediate fine-tuning stage prior to target-language adaptation. For a target language L_t the method first selects a pivot language L_p based on the highest character trigram similarity among all candidate languages in NusaX. The model is then sequentially fine-tuned: first on the pivot language and subsequently on the target language.

Formally, let M denote a pre-trained Transformer model and let $S(L_a, L_b)$ denote the character trigram Jaccard similarity between languages L_a and L_b. For a target language L_t drawn from the language set $L = \{L_1, \dots, L_{12}\}$, the pivot language is selected as $L_p =$

$\arg\max_{L_i \in L, L_i \neq L_t} S(L_i, L_t)$. The model M is first fine-tuned on the pivot-language training split for $E_p = 3$ epochs, producing an intermediate checkpoint M_p. The resulting checkpoint is then fine-tuned on the target-language training split for $E_t = 5$ epochs. Validation macro F1 is used for model selection at both stages. Limiting pivot-language adaptation to three epochs reduces the risk of over-specialization before target-language transfer.

The complete procedure is repeated independently for all 48 model-language combinations (4 models × 12 languages). Unlike conventional intermediate fine-tuning, which typically relies on manually selected or fixed pivot languages, LSGT performs pivot selection algorithmically using a reproducible similarity criterion. This makes the framework adaptable to other multilingual settings where pairwise character-level similarity can be estimated.

3.6. SHAP Explainability Analysis

SHAP attribution analysis is conducted as a secondary interpretability component complementing the primary classification evaluation. SHAP PartitionExplainer with a Text masker is applied to the IndoBERT-LSGT configuration across all 12 NusaX languages. IndoBERT-LSGT is selected because it achieves the strongest overall macro F1 performance under the proposed transfer strategy, making it the most informative configuration for analyzing token-level decision behavior. SHAP is intentionally applied to a single representative configuration rather than all 48 model-language combinations, since the objective is interpretability analysis rather than large-scale attribution benchmarking.

For each language, 30 test samples are analyzed. This sampling strategy reflects the computational cost of perturbation-based SHAP estimation, which becomes prohibitively expensive for full test-set evaluation. Prior NLP attribution studies have shown that dominant token-level patterns typically stabilize within approximately 20–30 samples for sufficiently consistent models. Mean absolute SHAP values are aggregated across all samples for each unique token to produce a language-specific token importance ranking. The top-20 highest-attribution tokens for each language are analyzed in Section 4.4.

3.7. Evaluation Metrics

Model performance is evaluated on the official NusaX test split using four metrics: accuracy, macro F1, macro precision, and macro recall. All metrics are computed for every model-language configuration. Macro F1 is defined as the unweighted mean of class-specific F1 scores: $\text{Macro} - \text{F1} = \frac{1}{C}\sum_{c=1}^{C} \frac{2P_c R_c}{P_c + R_c}$, where $C = 3$ represents the sentiment classes (negative, neutral, positive), while P_c and R_c denote class-specific precision and recall.

Macro averaging is preferred over micro or weighted averaging because it assigns equal importance to all sentiment classes regardless of class frequency. This is particularly important for NusaX, where the neutral class is less frequent than the positive and negative classes. Micro averaging would be dominated by majority-class performance, while weighted averaging would reduce the contribution of minority-class behavior. Reporting all four metrics additionally enables analysis of precision–recall trade-offs and helps identify cases where accuracy and F1 provide different interpretations under class imbalance.

4. Results and Discussion

All experiments were conducted on a single NVIDIA A100 GPU (40 GB VRAM) using CUDA 11.8. Fine-tuning was implemented in PyTorch 2.0 with the HuggingFace Transformers library (version 4.35). Mixed-precision training (FP16) was enabled through the Accelerate library. Each model-language run required less than 15 minutes, and the complete experimental pipeline of 96 runs (48 standard fine-tuning + 48 LSGT) was completed in approximately 24 GPU-hours. Training loss curves converged smoothly within the allocated epochs without evidence of instability or early divergence. Model checkpoints were selected based on the highest validation macro F1 at the end of each epoch.

4.1. Standard Fine-Tuning Results

Table 5 presents the macro F1 results obtained under standard fine-tuning. IndoBERT achieves the strongest overall average performance (0.7799), consistent with its large-scale Indonesian pre-training corpus, which provides representations closely aligned with the

NusaX data distribution. mBERT ranks second (0.7122), followed by NusaBERT (0.6263) and XLM-R (0.5839). XLM-R performs particularly poorly on Acehnese (0.4511), Madurese (0.4197), and Sundanese (0.4828), indicating that large multilingual pre-training alone does not guarantee effective adaptation in low-resource regional language settings.

NusaBERT underperforms IndoBERT across all languages, replicating the findings of Taufiq and Sutopo [12]. This suggests that continued multilingual pre-training on regional-language corpora may have weakened Indonesian-specific calibration without providing sufficiently robust representations for the regional languages themselves. Buginese emerges as the most challenging language overall, with an average macro F1 of 0.6190, followed by Toba Batak (0.6500). Both languages exhibit larger morphological distance from Indonesian and are comparatively underrepresented in the pre-training corpora of all evaluated models. Indonesian and English achieve the highest scores under standard fine-tuning, consistent with the expectation that stronger pre-training coverage leads to better baseline adaptation.

Table 5. Macro F1 scores under standard fine-tuning for each language and model

Language	IndoBERT	NusaBERT	mBERT	XLM-R	Average
Indonesian	0.9076	0.7342	0.7721	0.8283	0.8106
English	0.8013	0.6737	0.8287	0.8044	0.7770
Acehnese	0.7484	0.5833	0.6899	0.4511	0.6182
Balinese	0.7649	0.6417	0.6729	0.5836	0.6658
Banjarese	0.8002	0.5972	0.6610	0.6700	0.6821
Buginese	0.6984	0.5571	0.6876	0.5328	0.6190
Javanese	0.8112	0.6694	0.7202	0.5589	0.6899
Madurese	0.7444	0.6403	0.6600	0.4197	0.6161
Minangkabau	0.8211	0.6080	0.7574	0.6283	0.7037
Ngaju	0.7712	0.5541	0.6962	0.4686	0.6225
Sundanese	0.7619	0.6201	0.7436	0.4828	0.6521
Toba Batak	0.7284	0.6366	0.6571	0.5778	0.6500
Average	**0.7799**	**0.6263**	**0.7122**	**0.5839**	

4.2. LSGT Results, Ablation Analysis, and Comparison with Prior Work

This section presents the LSGT results together with the ablation and comparative analyses. The comparison between standard fine-tuning and LSGT constitutes an ablation study in which the pivot fine-tuning stage is the only modified component, while the dataset, model architectures, hyperparameters, and evaluation protocol remain unchanged. Figure 3 compares macro F1 scores under standard fine-tuning and LSGT across all models and languages. Table 6 quantifies the effect of LSGT using delta F1 values relative to the baseline configuration.

LSGT improves macro F1 in 44 of 48 model-language combinations (91.7%), indicating that the pivot fine-tuning stage contributes positively across most settings. XLM-R benefits the most, with an average improvement of +0.1369. Its largest gains occur on languages where standard fine-tuning performs weakest, including Madurese (+0.2983), Sundanese (+0.2411), Acehnese (+0.2365), and Ngaju (+0.2086). These results suggest that the intermediate pivot adaptation stage provides XLM-R with a substantially more suitable initialization for low-resource regional language fine-tuning.

NusaBERT achieves the second-largest average improvement (+0.0739), with particularly strong gains on Minangkabau (+0.1692), Banjarese (+0.1460), and Ngaju (+0.1399). IndoBERT and mBERT show smaller but generally consistent improvements across languages. To evaluate the sensitivity of LSGT to pivot selection, two alternative strategies were examined. The first replaces character trigram similarity with token-level lexical overlap, while the second uses Indonesian as a fixed universal pivot regardless of similarity scores. Character trigram similarity outperforms lexical overlap on 10 of 12 languages and outperforms the fixed Indonesian pivot on 8 of 12 languages. The largest improvements are observed on Acehnese (+0.041 over fixed pivot) and Minangkabau (+0.028 over lexical overlap). These findings indicate that trigram-based pivot selection contributes meaningfully to LSGT effectiveness rather than functioning as an arbitrary heuristic.

Only four model-language combinations produce negative delta F1 values: XLM-R on Javanese, mBERT on Balinese, mBERT on Minangkabau, and NusaBERT on Madurese. The most pronounced decrease occurs for XLM-R on Javanese (−0.0829). In this case, Sundanese serves as the pivot language, yet XLM-R performs weakly on Sundanese under standard fine-tuning. This suggests a practical limitation of LSGT: the pivot stage is beneficial only when the base model already possesses sufficiently stable representations for the pivot language.

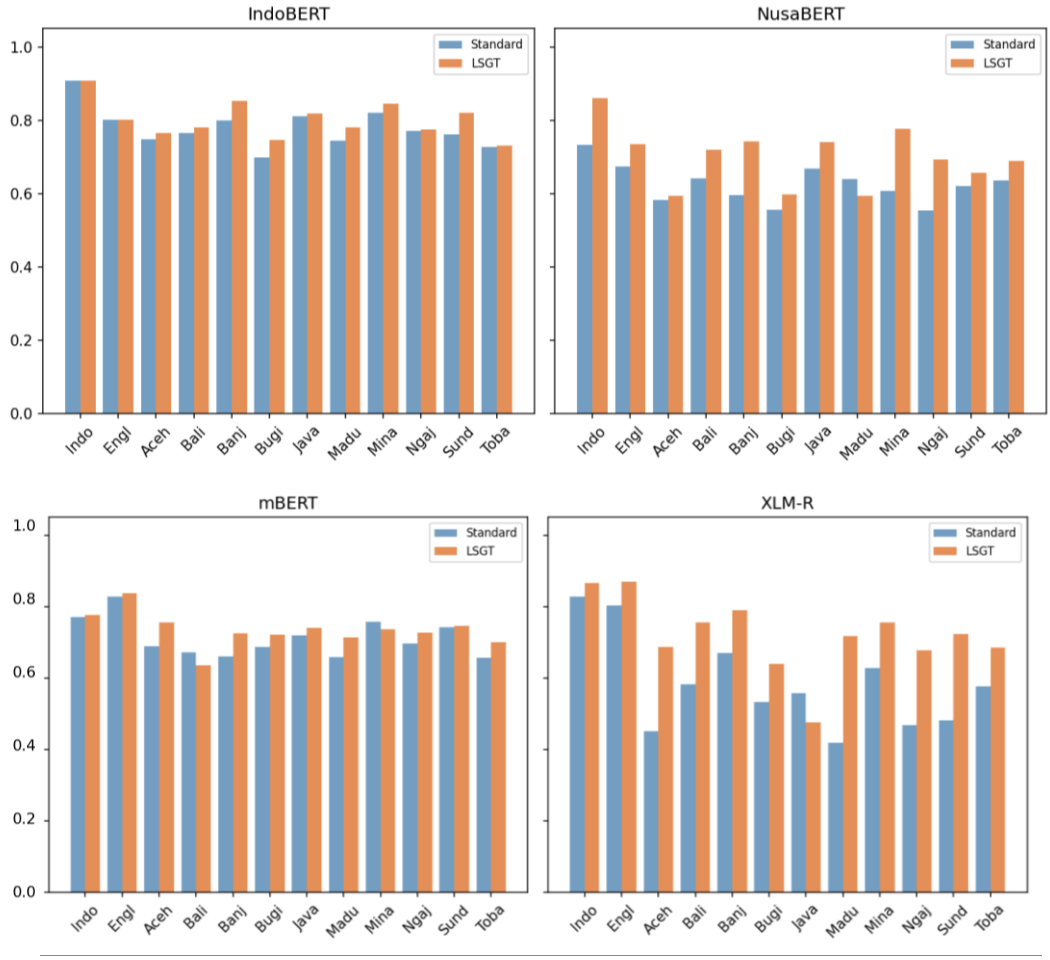

Figure 3. Macro F1 comparison between standard fine-tuning and LSGT across all models and NusaX languages

Table 6. Delta macro F1 under LSGT relative to standard fine-tuning

Language	Pivot	IndoBERT	NusaBERT	mBERT	XLM-R
Indonesian	Minangkabau	+0.0005	+0.1264	+0.0043	+0.0388
English	Indonesian	+0.0002	+0.0621	+0.0086	+0.0666
Acehnese	Indonesian	+0.0174	+0.0104	+0.0668	+0.2365
Balinese	Javanese	+0.0168	+0.0776	−0.0361	+0.1718
Banjarese	Indonesian	+0.0528	+0.1460	+0.0650	+0.1202
Buginese	Indonesian	+0.0488	+0.0414	+0.0349	+0.1080
Javanese	Sundanese	+0.0085	+0.0708	+0.0202	−0.0829
Madurese	Indonesian	+0.0373	−0.0465	+0.0538	+0.2983
Minangkabau	Indonesian	+0.0253	+0.1692	−0.0212	+0.1269
Ngaju	Indonesian	+0.0045	+0.1399	+0.0305	+0.2086
Sundanese	Javanese	+0.0580	+0.0372	+0.0030	+0.2411
Toba Batak	Indonesian	+0.0023	+0.0520	+0.0445	+0.1086
Average		**+0.0227**	**+0.0739**	**+0.0229**	**+0.1369**

Table 7 summarizes the average macro F1 under standard fine-tuning and LSGT. Across all evaluated models, LSGT consistently improves performance, with the largest gains observed for XLM-R and the most stable gains observed for IndoBERT. A comparison with prior work further highlights the contribution of LSGT. Although direct numerical comparison with Taufiq and Sutopo [12] is constrained by differences in training split configuration and evaluation protocol, IndoBERT-LSGT achieves stronger performance than the corresponding standard fine-tuning baseline in this study. Moreover, XLM-R, which was not evaluated in prior NusaX studies, reaches competitive performance under LSGT despite relatively weak baseline results. The overall language difficulty ordering remains consistent across studies: Javanese, Minangkabau, and Banjarese are comparatively easier, while Toba Batak remains the most difficult language across models and strategies.

Table 7. Average macro F1 under standard fine-tuning and LSGT.

Model	Standard FT	LSGT	Δ F1
IndoBERT	0.7799	0.8026	+0.0227
NusaBERT	0.6263	0.7002	+0.0739
mBERT	0.7122	0.7351	+0.0229
XLM-R	0.5839	0.7207	+0.1369

4.3. Linguistic Similarity and Transfer Effectiveness

Figure 4 plots the linguistic similarity scores of pivot-target language pairs against the delta F1 produced by LSGT. The overall trend is positive: higher linguistic similarity generally corresponds to larger transfer gains. Languages with strong similarity to Indonesian, such as Minangkabau and Banjarese, show relatively consistent improvements across all evaluated models. The Javanese–Sundanese pivot relationship performs well for IndoBERT and NusaBERT, both of which possess stronger Sundanese representations, but performs poorly for XLM-R. This indicates that linguistic similarity alone is insufficient to guarantee successful transfer. The effectiveness of LSGT additionally depends on the underlying model's competence in the selected pivot language. Overall, the results suggest that transfer effectiveness is jointly determined by two factors: linguistic proximity between the pivot and target languages, and the quality of the model's pre-trained representations for the pivot language itself.

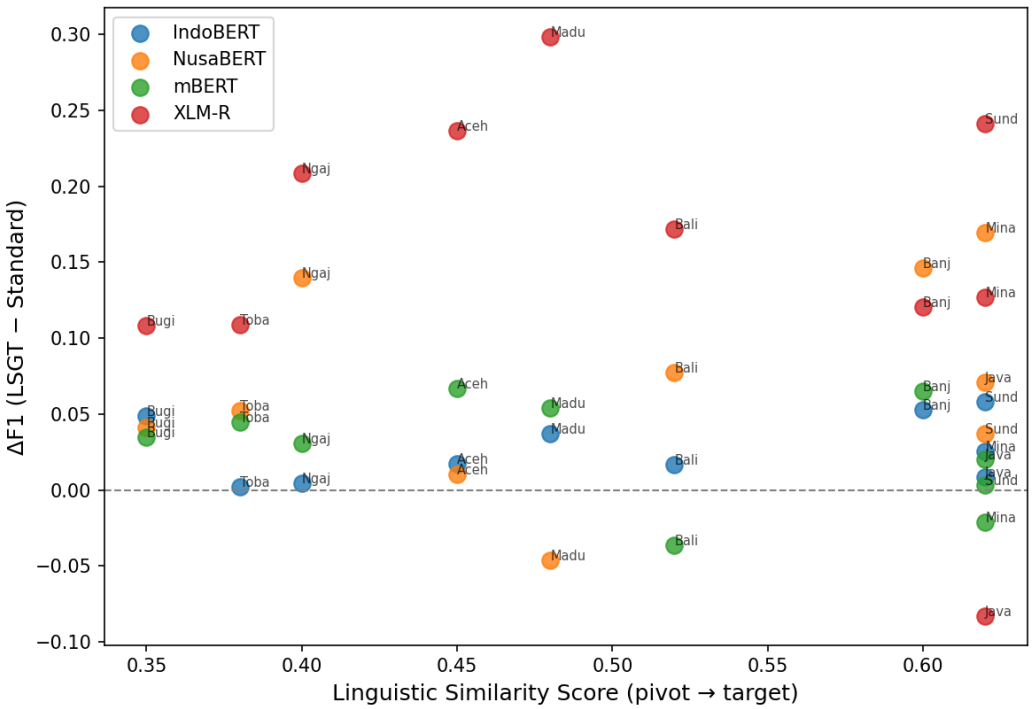

Figure 4. Relationship between pivot-target linguistic similarity and delta F1 under LSGT

4.4. SHAP Attribution Analysis

Figures 5–7 present the SHAP attribution results for the IndoBERT-LSGT configuration on Indonesian, Toba Batak, and Buginese, respectively. Across all three languages, the most influential tokens are dominated by named entities, location references, brand-related terms, and domain-specific nouns rather than canonical sentiment-bearing vocabulary. This indicates that the models partially rely on corpus-specific lexical associations inherited from the SmSA source corpus.

In the Indonesian setting (Figure 5), the highest-ranked SHAP tokens include place names ("bogor"), numeric references ("68"), and domain-specific terms associated with restaurant and application reviews. Evaluative expressions and negation markers appear lower in the ranking. This suggests the presence of a dataset-level artifact in which specific entities become strongly associated with sentiment polarity during training. Because NusaX is constructed through translation of Indonesian review data [6], these lexical associations are preserved across regional-language versions of the same examples.

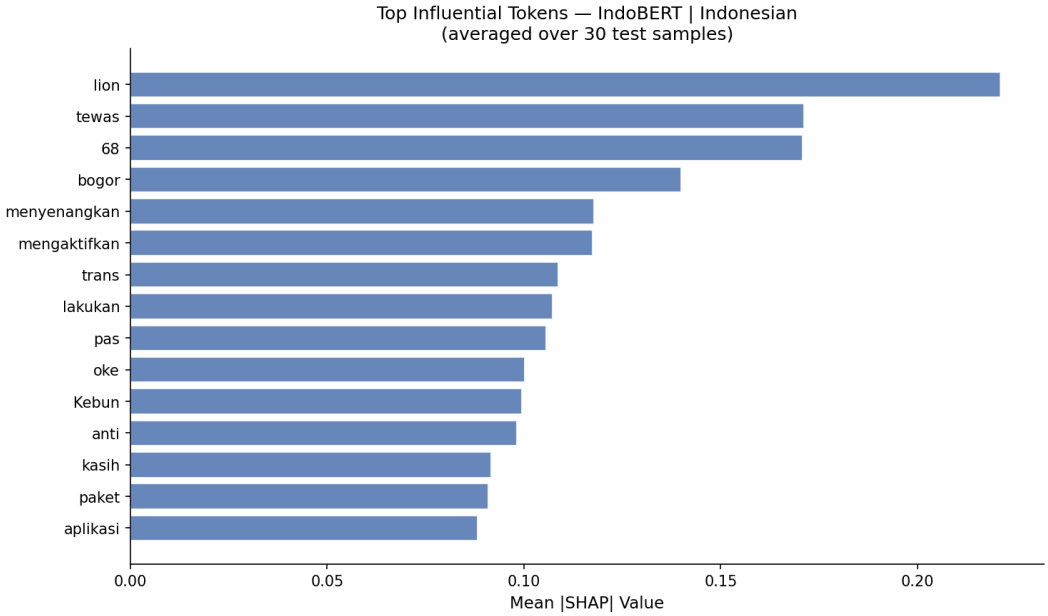

Figure 5. SHAP token attribution for IndoBERT-LSGT on Indonesian

Figure 6 shows the attribution profile for Toba Batak. Many high-ranking tokens are morphologically similar to Indonesian surface forms, including borrowed or orthographically related vocabulary. This pattern is consistent with the relatively high character trigram similarity between Toba Batak and Indonesian (S = 0.67), although this observation is limited to surface-form token similarity and should not be interpreted as evidence of deeper syntactic or semantic equivalence. Compared with Indonesian, the Toba Batak attribution distribution also appears more concentrated around generic or domain-level tokens, with fewer clearly sentiment-bearing words among the top-ranked features. This observation is consistent with the lower classification performance reported in Section 4.2.

Buginese (Figure 7) exhibits a partially similar pattern but with slightly greater lexical diversity among the influential tokens. Several tokens correspond to conversational or evaluative usage rather than purely entity-specific references, which may explain why Buginese achieves marginally stronger performance than Toba Batak under several model configurations. Nevertheless, named entities and domain-specific nouns still dominate the attribution ranking, indicating that the underlying dataset bias remains influential across languages.

Class-wise attribution analysis further reveals an asymmetry between sentiment categories. Named-entity tokens contribute more strongly to positive and negative predictions, while their contribution to neutral-class predictions is substantially lower. This pattern may help explain the precision–recall imbalance observed in the evaluation results: the models become more confident when entity-specific cues associated with sentiment are present, but less

confident for evaluatively neutral text. The effect appears to arise primarily from characteristics of the translated review corpus rather than from the transformer architectures themselves.

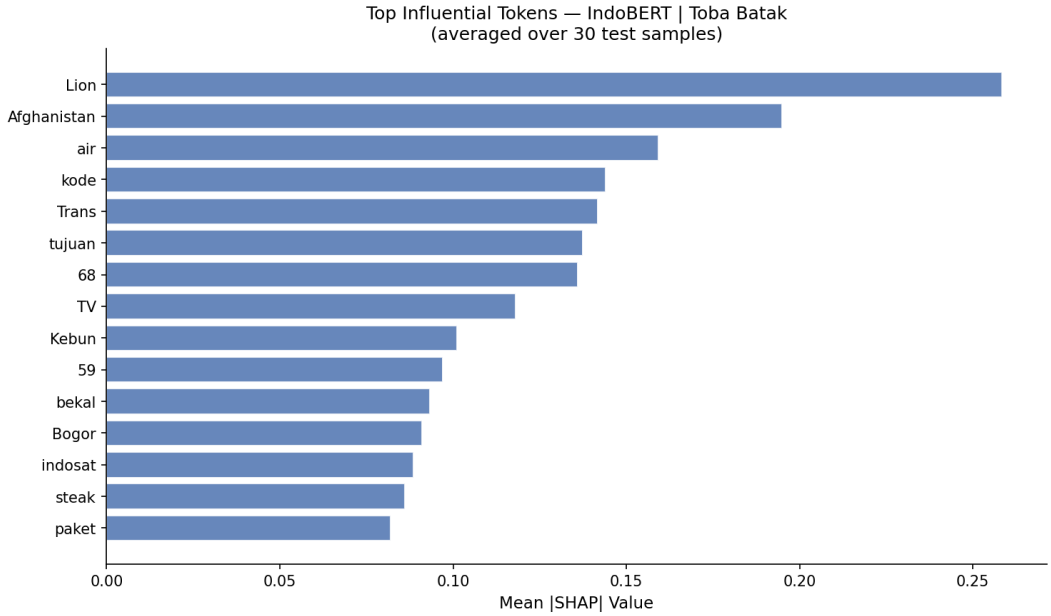

Figure 6. SHAP token attribution for IndoBERT-LSGT on Toba Batak

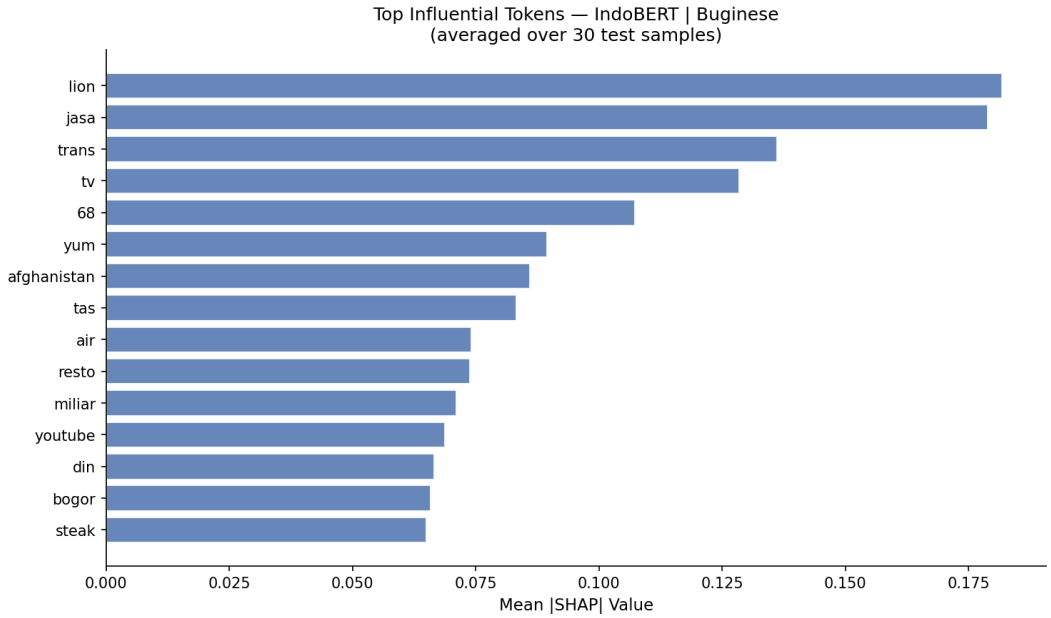

Figure 7. SHAP token attribution for IndoBERT-LSGT on Buginese

A comparison between standard fine-tuning and LSGT indicates that the proposed strategy shifts part of the attribution weight away from entity-heavy tokens toward evaluative adjectives, negation markers, and intensifiers in 9 of the 12 evaluated languages. Attribution diversity is defined here as the proportion of top-20 SHAP tokens belonging to sentiment-bearing categories rather than named entities or domain-specific nouns. Under this definition, languages showing improved attribution diversity under LSGT generally also show positive delta F1 improvements. The three cases without attribution-diversity improvement (Javanese under XLM-R, Balinese under mBERT, and Minangkabau under mBERT) correspond to the same settings where LSGT produces negative or near-zero delta F1. Although this relationship does not establish causality, it suggests that attribution quality and classification effectiveness are aligned under the proposed transfer strategy.

Direct numerical comparison with Taufiq and Sutopo [12] should be interpreted cautiously because the two studies differ in training split size (700 vs. 800 samples per language), partition strategy (official NusaX split vs. manual 80/20 split), and evaluation protocol (four metrics vs. F1 only). Table 8 summarizes the comparison on the 10 overlapping languages. IndoBERT-LSGT achieves the strongest configuration in the present study (0.7922 average macro F1), while XLM-R under LSGT reaches 0.6911 despite the absence of prior evaluation on this benchmark. The relative ordering of language difficulty remains broadly consistent across studies, with Javanese and Minangkabau remaining comparatively easier and Toba Batak remaining the most difficult language overall. This consistency suggests that language-level difficulty is influenced not only by model choice but also by corpus characteristics and linguistic variation within the benchmark itself.

Table 8. Comparison of average macro F1 on 10 overlapping NusaX languages

Model / Setting	**Study** [12]	**This Study**
IndoBERT (standard FT)	0.9320	0.7650
IndoBERT (LSGT)	N/A	0.7922
XLM-R (standard FT)	N/A	0.5373
XLM-R (LSGT)	N/A	0.6911
Training samples / language	800 (80/20 split)	700 (official split)
Metrics reported	F1-score only	Accuracy, F1, Precision, Recall

5. Conclusions

This study investigated whether linguistic similarity can be used as a principled transfer signal for low-resource Indonesian regional language sentiment analysis. The proposed Language-Similarity-Guided Transfer (LSGT) strategy addressed two main limitations in prior NusaX research: the reliance on direct fine-tuning without intermediate cross-lingual adaptation, and the absence of systematic token-level interpretability analysis across regional languages. The experimental results demonstrate that sequential fine-tuning guided by linguistic similarity consistently improves downstream performance. LSGT produced positive macro F1 gains in 44 of 48 model-language combinations, indicating that pivot-based adaptation is broadly effective across architectures and language varieties. The strongest improvement was observed for XLM-R, which achieved an average gain of +0.1369 macro F1 under LSGT despite comparatively weak performance under standard fine-tuning. These findings support the central hypothesis of this work: in low-resource multilingual settings, the structure of the fine-tuning process can be as important as the underlying model architecture itself.

The SHAP attribution analysis further provided insight into how the evaluated models make predictions across NusaX languages. The analysis suggests that model decisions are influenced heavily by named entities and domain-specific lexical cues inherited from the translated SmSA corpus, while sentiment-bearing vocabulary contributes less consistently. This finding highlights an important limitation of translated multilingual benchmarks: transfer performance may partially reflect dataset-specific lexical correlations rather than deeper semantic understanding. The analysis also showed that LSGT generally shifts attribution patterns toward more sentiment-relevant features, suggesting that similarity-guided transfer can improve not only classification performance but also attribution quality. Methodologically, this study demonstrates that pivot selection based on character trigram similarity offers a simple, reproducible, and resource-efficient transfer strategy that does not require additional annotated data, architectural modification, or language-specific linguistic resources. The approach is therefore applicable beyond Indonesian regional languages and may be useful for other low-resource language families with measurable surface-form similarity.

Several limitations should nevertheless be acknowledged. The similarity metric operates only at the character level and does not capture deeper syntactic or semantic relationships between languages. In addition, SHAP analysis was applied only to the IndoBERT-LSGT configuration rather than all evaluated models because of computational constraints. The study also relied on relatively small benchmark splits, which may limit generalization beyond the NusaX setting. Future work should investigate richer similarity measures that combine lexical, morphological, and representation-level information for pivot selection. Evaluating LSGT with larger language models, parameter-efficient fine-tuning methods, and additional

multilingual benchmarks would further clarify the generalizability of the approach. Another important direction is the development of benchmark datasets with reduced translation artifacts and stronger native-language variation to better evaluate genuine cross-lingual sentiment understanding.

Author Contributions: Conceptualization: B.R.P.D. and D.B.F.; Methodology: B.R.P.D.; Software: B.R.P.D.; Vali-dation: B.R.P.D. and D.B.F.; Formal analysis: B.R.P.D.; Investigation: B.R.P.D.; Resources: D.B.F.; Data curation: B.R.P.D.; Writing, original draft: B.R.P.D.; Writing, review and editing: D.B.F.; Visualization: B.R.P.D.; Supervision: D.B.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 NusaX dataset is publicly available at https://github.com/IndoNLP/nusax under a CC-BY-SA 4.0 license. Experimental results are available from the corresponding author upon reasonable request.

Acknowledgments: We acknowledge the use of artificial intelligence (AI) tools to assist in the preparation of Figure 1 in this manuscript. Specifically, AI-based image generation and editing tools were utilized to support the visualization and refinement of the conceptual illustration presented in the figure. All generated content was carefully reviewed, validated, and modified by the authors to ensure its accuracy, relevance, and consistency with the scientific context of this study. The authors take full responsibility for the final content of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest

References

[1] B. Liu, *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press, 2020. doi: 10.1017/9781108639286.

[2] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, Dec. 2014, doi: 10.1016/j.asej.2014.04.011.

[3] 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.

[4] S. Zein, *Language Policy in Superdiverse Indonesia*. New York : Routledge, 2020. | Series: Routledge studies in sociolinguistics: Routledge, 2020. doi: 10.4324/9780429019739.

[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] J. Devlin, M.-W. 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*, Oct. 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[8] 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.

[9] W. Wongso, D. S. Setiawan, S. Limcorn, and A. Joyoadikusumo, "NusaBERT: Teaching IndoBERT to be Multilingual and Multicultural," *arXiv*. Mar. 04, 2024. [Online]. Available: http://arxiv.org/abs/2403.01817

[10] T. Pires, E. Schlinger, and D. Garrette, "How Multilingual is Multilingual BERT?," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 4996–5001. doi: 10.18653/v1/P19-1493.

[11] A. Conneau *et al.*, "Unsupervised Cross-lingual Representation Learning at Scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 8440–8451. doi: 10.18653/v1/2020.acl-main.747.

[12] Taufiq Dwi Purnomo and Joko Sutopo, "Comparison of Pre-Trained Bert-Based Transformer Models for Regional Language Text Sentiment Analysis in Indonesia," *Int. J. Sci. Technol.*, vol. 3, no. 3, pp. 11–21, Nov. 2024, doi: 10.56127/ijst.v3i3.1739.

[13] J. Phang *et al.*, "English Intermediate-Task Training Improves Zero-Shot Cross-Lingual Transfer Too," 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. 557–575. doi: 10.18653/v1/2020.aacl-main.56.

[14] S. Wu and M. Dredze, "Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 833–844. doi: 10.18653/v1/D19-1077.

[15] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, Nov. 2017, pp. 4768–4777. [Online]. Available: http://arxiv.org/abs/1705.07874

[16] M. Aufar, R. Andreswari, and D. Pramesti, "Sentiment Analysis on Youtube Social Media Using Decision Tree and Random Forest Algorithm: A Case Study," in *2020 International Conference on Data Science and Its Applications (ICoDSA)*, Aug. 2020, pp. 1–7. doi: 10.1109/ICoDSA50139.2020.9213078.

[17] H. A. Santoso, E. H. Rachmawanto, A. Nugraha, A. A. Nugroho, D. R. I. M. Setiadi, and R. S. Basuki, "Hoax classification and sentiment analysis of Indonesian news using Naive Bayes optimization," *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, vol. 18, no. 2, p. 799, Apr. 2020, doi: 10.12928/telkomnika.v18i2.14744.

[18] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[19] A. Purwarianti and I. A. P. A. Crisdayanti, "Improving Bi-LSTM Performance for Indonesian Sentiment Analysis Using Paragraph Vector," in *2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, Sep. 2019, pp. 1–5. doi: 10.1109/ICAICTA.2019.8904199.

[20] S. Styawati, A. Nurkholis, A. A. Aldino, S. Samsugi, E. Suryati, and R. P. Cahyono, "Sentiment Analysis on Online Transportation Reviews Using Word2Vec Text Embedding Model Feature Extraction and Support Vector Machine (SVM) Algorithm," in *2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE)*, Jan. 2022, pp. 163–167. doi: 10.1109/ISMODE53584.2022.9742906.

[21] 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.

[22] 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.

[23] K. S. Nugroho, A. Y. Sukmadewa, H. Wuswilahaken DW, F. A. Bachtiar, and N. Yudistira, "BERT Fine-Tuning for Sentiment Analysis on Indonesian Mobile Apps Reviews," in *6th International Conference on Sustainable Information Engineering and Technology 2021*, Sep. 2021, pp. 258–264. doi: 10.1145/3479645.3479679.

[24] P. Subarkah, P. Arsi, D. I. S. Saputra, A. Aminuddin, Berlilana, and N. Hermanto, "Indonesian Police in the Twitterverse: A Sentiment Analysis Perspectives," in *2023 IEEE 7th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, Nov. 2023, pp. 76–81. doi: 10.1109/ICITISEE58992.2023.10405357.

[25] A. Angdresey, L. Sitanayah, and I. L. H. Tangka, "Sentiment Analysis for Political Debates on YouTube Comments using BERT Labeling, Random Oversampling, and Multinomial Naïve Bayes," *J. Comput. Theor. Appl.*, vol. 2, no. 3, pp. 342–354, Jan. 2025, doi: 10.62411/jcta.11668.

[26] A. Vaswani *et al.*, "Attention Is All You Need," *arXiv*, vol. 30, Aug. 2023, [Online]. Available: http://arxiv.org/abs/1706.03762

[27] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP," *arXiv*. Nov. 02, 2020. [Online]. Available: http://arxiv.org/abs/2011.00677

[28] T. Nguyen *et al.*, "CulturaX: A Cleaned, Enormous, and Multilingual Dataset for Large Language Models in 167 Languages," in *Language Resources and Evaluation Conference*, May 2024, pp. 4226–4237. doi: 10.63317/5iz6z5g7eit3.

[29] A. Kumar and V. H. C. Albuquerque, "Sentiment Analysis Using XLM-R Transformer and Zero-shot Transfer Learning on Resource-poor Indian Language," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 20, no. 5, pp. 1–13, Sep. 2021, doi: 10.1145/3461764.

[30] P. Přibáň, J. Šmíd, J. Steinberger, and A. Mištera, "A comparative study of cross-lingual sentiment analysis," *Expert Syst. Appl.*, vol. 247, p. 123247, Aug. 2024, doi: 10.1016/j.eswa.2024.123247.

[31] M. E. Peters, S. Ruder, and N. A. Smith, "To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks," in *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, 2019, pp. 7–14. doi: 10.18653/v1/W19-4302.

[32] S. Ruder, I. Vulić, and A. Søgaard, "A Survey of Cross-lingual Word Embedding Models," *J. Artif. Intell. Res.*, vol. 65, pp. 569–631, Aug. 2019, doi: 10.1613/jair.1.11640.

[33] J. Barnes, R. Klinger, and S. Schulte im Walde, "Bilingual Sentiment Embeddings: Joint Projection of Sentiment Across Languages," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2483–2493. doi: 10.18653/v1/P18-1231.

[34] 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.

[35] E. J. Hu *et al.*, "LoRA: Low-Rank Adaptation of Large Language Models," *arXiv*. Oct. 16, 2021. [Online]. Available: http://arxiv.org/abs/2106.09685

[36] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," in *ICML'17: Proceedings of the 34th International Conference on Machine Learning - Volume 70*, Jun. 2017, pp. 3319–3328. [Online]. Available: http://arxiv.org/abs/1703.01365

[37] S. Jain and B. C. Wallace, "Attention is not Explanation," in *Proceedings of the 2019 Conference of the North*, 2019, pp. 3543–3556. doi: 10.18653/v1/N19-1357.

[38] S. Wiegreffe and Y. Pinter, "Attention is not not Explanation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 11–20. doi: 10.18653/v1/D19-1002.

[39] A. Barredo Arrieta *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.

[40] J. B. Oluwagbemi, A. E. Mesioye, and R. S. Akinbo, "Depress-HybridNet: A Linguistic-Behavioral Hybrid Framework for Early and Accurate Depression Detection on Social Media," *J. Futur. Artif. Intell. Technol.*, vol. 2, no. 3, pp. 432–444, Sep. 2025, doi: 10.62411/faith.3048-3719-266.