JCBI
Journal of Computing & Biomedical Informatics Volume 10 Issue 02
ISSN: 2710 - 1606 2026

Research Article
https://doi.org/10.56979/1002/2026/1332

Analyzing the Impact of Pretrained Language Models on Low-Resource Languages

Muhammad Irshad Hussain[1], Shafiq Hussain[2*], Aleena Jamil[2], Adeen Amjad[2], and Sajid Iqbal[3]

[1]Riyadh Lake Real Estate Development Company (PIF subsidiary), Saudi Arabia.
[2]Department of Computer Science, University of Sahiwal, Sahiwal, 57000, Pakistan.
[3]Department of Information Systems, King Faisal University, Al-Ahsa, Saudi Arabia.
[*]Corresponding Author: Shafiq Hussain. Email: drshafiq@uosahiwal.edu.pk

Received: December 29, 2025 **Accepted:** February 28, 2026

Abstract: The rapid advancement of natural language processing has predominantly benefited high-resource languages such as English, Chinese, and Spanish, leaving thousands of languages underserved. This digital language divide limits equitable access to technology and threatens global linguistic diversity. This paper presents a systematic evaluation of eight pretrained language models across seven low-resource languages representing five distinct language families. Through extensive experiments on sentiment analysis, named entity recognition, and machine translation tasks, we demonstrate that multilingual BERT achieves the highest average accuracy of 74.5%. We further propose a novel adaptation framework combining vocabulary augmentation, continual pretraining, task-adaptive fine-tuning, and knowledge distillation that improves performance by up to 18.7%. Our analysis identifies vocabulary overlap as the strongest predictor of cross-lingual transfer success, explaining 76.3% of performance variance. These findings provide evidence-based guidelines for researchers and practitioners developing inclusive NLP technologies for underserved language communities. Limitations of this study include the focus on seven languages (generalizability to other low-resource languages requires further validation), computational constraints that prevented evaluation of models exceeding 300M parameters, and potential biases introduced by dataset availability and quality across languages.

Keywords: Low-Resource Languages; Pretrained Language Models; Cross-Lingual Transfer; Multilingual NLP; Adaptation Framework

1. Introduction

The field of Natural Language Processing (NLP) has witnessed an exponential rise in recent times, mainly due to the introduction of Pre-Trained Language Models (PLMs), like BERT [2] and GPT [3], alongside their multilingual versions. Such models have been shown to yield state-of-the-art results for diverse NLP tasks, ranging from text categorization and Named Entity Recognition (NER) to Machine Translation and Question Answering (QA). Nevertheless, all of this development has occurred mainly for a few well-resourced languages, mainly English, Chinese, and Spanish, leaving most other languages highly underrepresented. As pointed out by [4], such a technology gap has led to an unequal digital language gap [5].

Among 7,000 languages, fewer than 100 have adequate digital resources to enable contemporary natural language processing [6]. For fully resourced languages, there is the possibility of training materials measured in billions of language tokens, whereas for low-resourced languages, including indigenous and regional languages in Asia, Africa, and South America, there is no digital footprint at all [7]. The difference is not only a matter of lack of access to technology, but it also affects the diversity of language in the digital age [8].

The challenges in building proficient NLP solutions for low- resource languages are intricate and varied. These have been shortlisted as part of previous surveys in the field [9], [18] and include the following: (1) Lack of annotated data resources, (2) unavailability of monolingual corpora resources, (3) Lack of primary resources like tokenizing resources, parsing resources, and lexical resources, (4) high morphological complexity, and (5) script diversity. This results in the situation where it is difficult to construct proficient models using traditional data- driven techniques.

Cross-lingual transfer learning in multilingual pre-trained language models (PLMs) is a promising direction in overcoming these issues. Basic research on models like multilingual BERT (mBERT) [22], and XLM-R [27] has shown that it is possible to transfer the learned representations from high- resource to low-resource languages through shared embedding spaces. However, the degree of transferability is inconsistent; research studies such as [32] and others [25] have shown that typological similarity, lexical overlap, and the composition of pre-training data have a significant impact on cross-lingual transferability. Thus, the current state of knowledge on the performance of different PLM architectures across low-resource language scenarios is incomplete.

To address these gaps, we conduct a systematic investigation centered on the following research questions:

1) **RQ1:** How do various architectures of pretrained lan- guage models (PLMs) compare on low-resource lan- guages to their performance on high-resource languages?
2) **RQ2:** What are the most influential linguistic, architectural, and data-related factors for PLM performance in low-resource environments?
3) **RQ3:** What are the most important architectural and pretraining factors that have a significant impact on the effectiveness of cross-lingual transfer to low-resource languages?
4) **RQ4:** Can systematic adaptation and fine-tuning strategies effectively improve PLM performance on low- resource languages, and if so, which strategies are most effective?
5) **RQ5:** What evidence-based guidelines can be developed for the selection, adaptation, and evaluation of PLMs in low-resource environments?

To answer these questions, we make the following key contributions:

Comprehensive Empirical Comparison: We conduct the most extensive empirical evaluation to date of eight state-of-the-art PLMs across five low-resource language families (Tamil, Urdu, Filipino, Swahili, Uzbek, Bengali, Yoruba) on multiple NLP tasks, supported by rigorous statistical significance testing.

Novel Adaptation Framework: We propose and evaluate a new adaptation strategy that effectively integrates vocabulary expansion, pretraining, fine-tuning, and knowledge distillation, which brings performance gains of up to 18.7% relative to fine-tuning.

Evidence-Based Guidelines: We provide practical and data-driven insights into model selection, model adaptation, data usage, and evaluation in low-**resource** settings.

The rest of this paper is structured as follows: In Section II, we discuss existing literature relevant to low-resource NLP and multilingual models. Section III covers theoretical background and definitions. Our experiments' methods, including datasets and models used and their evaluation metrics, are described in Section IV. Our findings, along with their analysis, are presented in Section V, using appropriate statistical techniques.

2. Related Work

2.1. The *Low-Resource NLP Challenges*

The problem of low resources in NLP has been well reported in the literature over the past few years. [18] has a complete list of problems reported as data unavailability as a chief barrier, dividing it into four categories: (1) data unavailability, (2) language diversity, (3) evaluation practices, and (4) computational issues. Another paper, [26], analyzes these problems in the context of Indian languages and proves that a language can still be low-resource despite a large number of speakers.

More recent research has identified specific challenges: [27] investigates the rise of cross-lingual architectures in PLMs and examines the impact of model design on multilingual skills. The most effective methods for neural machine translation in low-resource languages have been investigated by [33]. For an application task, the issue of low-resource social media comment analysis for opinion-sentiment analysis

has been explored by [20], who proposed using LSTM networks and data augmentation methods to effectively address the problem despite limited training data.

2.2. Multilingual Pretrained Models

Multilingual PLMs have ushered in a paradigm shift in cross-lingual NLP. This breakthrough was made possible by the pioneering work in [2], which introduced multilingual BERT (mBERT), which demonstrated incredible zero-shot transfer performance on cross-lingual tasks. This was further enhanced in the study by [22].

Following these achievements, some architectures have emerged: mT5 is a massively multilingual T5, scaled to 101 languages, proposed in [31], while in [23], the architecture XNLI was introduced, a benchmark designed to assess cross-lingual sentence representation quality. Finally, in [24], better approaches to cross-lingual alignment were proposed, allowing zero

2.3. Cross-lingual Transfer Mechanisms

Studies on cross-lingual transfer mechanisms have recently attracted considerable interest. Pioneering studies by [32] showed that the effectiveness of cross-lingual transfer is significantly influenced by typological and lexical overlap between languages. The studies by [27] and [29] examine structural characteristics important for PLMs and relate them to the syntactic and semantic representations that PLMs learn before task-specific training. Theoretically, recent work by [25] formalized the notion of the cross-lingual transferability of monolingual representations, which showed the benefit of appropriate initialization. [28] Gave empirical results of the strong relationship between the efficiency of the transfer method and the linguistic similarity.

2.4. Adaptation Techniques for Low-Resource Languages

A number of adaptation strategies have also been suggested to improve PLM performance on low-resource languages. [17] Investigates methods to extend multilingual BERT to new languages by expanding the vocabulary and pretraining. [11] Introduces test-time adapter ensembling to efficiently adapt to low-resource language varieties.

Some recent developments in this area include [13], which reviews unsupervised learning methods for masked language models in low-resource languages and highlights the importance of transfer learning and data augmentation. [14], which presents an extensive description of transfer learning models in languages and their use cases.

1) *Tokenization and Morphological Challenges:* Tokenization is particularly important for high-morphological- complexity, low-resource languages. A thorough analysis of the impact of tokenization is provided by [15], which indicates that tokenization's subword strategies require careful optimization for agglutinative languages. The impact of the above factors on the Indian language is explored by [21].
2) In addition to tokenization, morphological complexity raises other research questions, as probed by [16], which explores transformer models for punctuation restoration in both resource-rich and resource-poor languages and reveals that knowledge of morphology can improve performance across script types. Machine translation performance for low-resource language varieties is also foregrounded by [12], which high- lights the need for appropriate tokenization strategies.

Evaluation Methodologies and Benchmarks

In fact, the construction of appropriate evaluation methodologies remains an active area of research work for low- resource languages. XNLI by [23] is a benchmark that standardizes evaluation across 15 languages. [19] surveyed neural machine translation evaluation, focusing on low-resource languages; this paper underscores that traditional accuracy measures should be complemented by a diverse set of evaluation metrics.

Recent Advances (2023-2025): Several recent works have advanced low-resource NLP beyond the studies cited above. Azhar et al. [34] proposed selective attention pruning for Urdu text summarization, achieving 15% inference speedup with minimal accuracy loss. Zhang et al. [35] introduced XLM-V, a multilingual model with a 256k vocabulary covering 100+ languages, showing particular improvements for morphologically rich languages. Adelani et al. [36] released MasakhaNEWS, a benchmark for 16 African languages, demonstrating that language-specific adaptations outperform generic multilingual

models. Furthermore, the emergence of BLOOM [37] (46 languages, 176B parameters) and LLaMA-2 multilingual variants [38] has expanded coverage for previously underserved languages, though their performance on very low-resource languages (corpora < 1M tokens) remains limited.

2.5. Background And Preliminaries

A. Definition and Taxonomy of Low-Resource Languages

We define a low-resource language, following [18] and [34], as one that lacks sufficient digital resources to develop standard statistical NLP systems. It also meets the definition in [26] for characterizing Indian languages and in [20] for performing social media analysis in low-resource settings.

We propose a multi-dimensional consideration:

1. Data Availability: As characterized by [13] and [19]

· *Critical*: Monolingual corpora (¡1M tokens), parallel corpora (¡10K pairs of sentences)
· *Severe*: 1-10M tokens of monolingual data with limited parallel data
· *Moderate*: Some parallel resources (10-100M tokens), monolingual corpora (10-100M tokens)

2. Linguistic Resources:

The analyses described in [5] and [11] have been completed.

· A lack of morphological analyzers, part-of-speech tags, and dependency parsers
· No named entity recognition and no sentiment lexicons
· Documentation of linguistics is lacking.

3. Digital Infrastructure: As discussed in [14] and [16]

· Limited web presence and digital content
· Sparse representation in social media platforms
· Inadequate keyboard/input method support

B. Formal Framework for Pretrained Language Models

We provide formal definitions and mathematical formula-tions for the key PLMs evaluated in our study, building upon the foundations established in [2], [22], and [27].

BERT and Variants: BERT [2] uses a transformer encoder architecture with the following formal specification, extending the analysis in [14]:

Architecture: Given input sequence $X = [x_1, x_2, ..., x_n]$, BERT computes contextual representations $H = [h_1, h_2, ..., h_n]$ through L transformer layers:

$H^0 = \text{Embedding}(X) + \text{PositionalEncoding}(X)$

$H^l = \text{TransformerLayer}(H^{l-1}),\ l = 1, ..., L$

Pretraining Objectives: As analyzed in [29]

1. Masked Language Modeling (MLM):

$$\boldsymbol{L_{MLM}} = E_{X\sim D}[\sum_{i\in M} -\log P(x_i | X\backslash_M)] \quad (1)$$

where M is a random subset of positions (typically 15%) masked during training.

2. Next Sentence Prediction (NSP):

$L_{NSP} = E_{(A,B)\sim D}[-\log P(\text{IsNext} | A, B)]$

3. Multilingual Extension (mBERT): Following [27], mBERT extends this framework to K languages by:

$$D_{multilingual} = \bigcup_{k=1}^{K} D_K \quad (2)$$

with balanced sampling across languages to prevent dominance by high-resource languages.

XLM-RoBERTa: XLM-R [22] extends RoBERTa to multilingual settings with optimizations analyzed in [24]:

4. Training Objective:

$$LXLM - R = k = 1\sum KE_X \sim D_k[i \in Mk\sum - logP(xi \mid XMk)] \quad (3)$$

5. *Cross-lingual Transfer Mechanisms:* Building on [25] and [28], we formalize cross-lingual transfer using information-theoretic principles:
6. Transfer Efficiency: Given source language S and target language T , transfer efficiency is defined as:

$$\eta S \to T = \frac{H(T)}{(T;\Theta S)} \quad (4)$$

where $I(T; \Theta_S)$ is the mutual information between target language data and source language parameters, and $H(T)$ is the entropy of the target language.

7. Vocabulary Overlap Metric: Extending [15] and [21], we define vocabulary overlap between languages i and j as:

$$Vij = min(|Vi|, |Vj|) \mid Vi \cap Vj \mid \quad (5)$$

where V_i is the vocabulary of language *i after tokenization*.

C. Adaptation Framework Theory

Our adaptation framework builds upon techniques surveyed in [11], [13], and [17]:

Given: Pretrained model M with parameters Θ, target language corpus D_T

Objective: Learn adapted parameters Θ' maximizing:

$$L(\Theta')=E(x,y)\sim DT[\log P(y|x;\Theta')]-\lambda DKL(P(\cdot|\Theta')\|P(\cdot|\Theta)) \quad (6)$$

where the KL-divergence term prevents catastrophic forgetting of multilingual knowledge.

D. Evaluation Metrics Formalism

Extending benchmarks discussed in [23] and [19], we define comprehensive evaluation metrics:

Transfer Efficiency Ratio (TER): As used in [27] and [28]

$$TER = \frac{Perf_{high-resource}}{Perf_{low-resource}} \quad (7)$$

Multilingual Utility Score (MUS): Incorporating insights from [14]

MUS = $\alpha \cdot$ Accuracy + $\beta \cdot$ TER + $\gamma \cdot (1 - \text{Params}/\text{Params}_{max})$ with $\alpha + \beta + \gamma = 1$ representing importance weights.

While our primary evaluation focuses on the eight models in Table II, we also benchmark against three recent multilingual models to contextualize our results: (1) BLOOM-176B [37], a 176-billion parameter decoder-only model covering 46 languages; (2) LLaMA-2-13B-multilingual [38], fine-tuned from LLaMA-2 on 20 languages; and a 95M parameter model pretrained exclusively on 11 African languages. These models were evaluated on a subset of tasks (sentiment analysis for Yoruba and Swahili, classification for Urdu) due to computational constraints (full evaluation would require >2000 GPU hours). BLOOM achieved 68.2% accuracy on Yoruba (vs. our adapted mBERT at 82.3%), LLaMA-2 achieved 71.4% on Swahili (vs. 79.2%), and AfriBERTa achieved 74.1% on Yoruba (vs. 82.3%). The specialized AfriBERTa outperforms generic multilingual models but still falls short of our adapted mBERT, highlighting the value of task-specific adaptation over domain-specific pretraining alone.

3. Methodology

3.1. This Experimental Design

The proposed study will make use of an elaborate experimental design for the evaluation of pretrained language models (PLMs) on low-resource languages. The research will respond to the first research question by: (1) elaborately designing the selection of the models based on various architecture designs, and (2) an elaborate design of the experiment's protocol evaluation that includes multiple NLP tasks conducted on low-resource languages. This experiment design will be elaborated in Figure 1.

Preprocessing Details: For each language, the following preprocessing steps were applied: (1) Unicode normalization to NFC form, (2) removal of non-linguistic characters (emojis, URLs) while preserving punctuation, (3) language-specific tokenization using each model's native tokenizer, (4) lowercasing for Latin-script languages (Filipino, Swahili, Uzbek, Yoruba) but preserved case for Arabic-script (Urdu) and Brahmic-script (Tamil, Bengali) languages. For machine translation tasks, we applied byte-pair encoding (BPE) with a vocabulary size of 32,000 subword units. The train/dev/test split used stratified sampling to preserve class distribution, with ratios of 80/10/10 for all datasets except Yoruba (75/12.5/12.5) due to smaller sample size.

3.2. Datasets

We curated datasets across five low-resource language families representing diverse linguistic typologies. Table I details the dataset characteristics.

3.3. Models Evaluated

We evaluated eight state-of-the-art PLMs representing diverse architectural paradigms:

Figure 2 provides a comprehensive comparison of these models across architectural and efficiency dimensions.

3.4. Experimental Setup

Experiments were conducted using NVIDIA Tesla V100 GPUs with 32 GB memory. The codes were written for transformers using the library provided by Hugging Face. Hyperparameter configurations included a learning rate of 2e-5, batch size of 32, number of epochs equal to 10, and early stopping with patience of 3. The experiments were replicated 5 times using various random seeds, and the standard deviation was computed based on their means.

3.5. Evaluation

Justification for Metric Selection: Accuracy was chosen as the primary metric for sentiment analysis and text classification because these tasks have balanced class distributions (verified via class proportion analysis, all within 45-55% range). F1-score (macro-averaged) is reported for NER due to inherent class imbalance (entity tags appear less frequently than non-entity tags). BLEU-4 is used for machine translation following standard practice in low-resource MT literature [19], with statistical significance assessed via bootstrap resampling. Transfer Efficiency Ratio (TER) is proposed as a novel metric to quantify cross-lingual transfer, defined as the ratio of low-resource to high-resource performance. This metric controls for task difficulty and enables fair comparison across models with different absolute performance ceilings.

A thorough evaluation was carried out based on a variety of evaluation criteria:

- Accuracy: The main evaluation metric for classification.
- F1-Score: Used for imbalanced datasets.
- BLEU Score: Used for machine translation evaluation
- Cross-lingual Transfer Efficiency (TER)
- Statistical Significance: Paired t-tests

3.6. Novel Adapt

The proposed adaptation framework consists of four components, each addressing a specific limitation of PLMs for low-resource languages:

Vocabulary Augmentation: We extend the model's tokenizer by adding 5,000-10,000 target-language-specific subwords (depending on language morphological complexity). This addresses the Out-Of-Vocabulary (OOV) problem, which we measured at 18-35% for English-only models on low-resource languages (see Table VIII). New subwords are identified using BPE on the target corpus and initialized using the average embeddings of semantically similar existing subwords.

Continual Pretraining: The augmented model undergoes additional masked language modeling pretraining on 100,000 target-language sentences (or full corpus if smaller). This reduces domain shift between the original pretraining distribution (primarily high-resource languages) and the target low-resource language. We use a learning rate of 5e-5 for 5 epochs.

Task-Adaptive Fine-Tuning (TAFT): Unlike standard fine-tuning, TAFT incorporates a KL-divergence regularization term to prevent catastrophic forgetting of multilingual knowledge. The loss function is $L_{total} = L_{task} + \lambda \cdot D_{KL}(P(\cdot|\Theta') \parallel P(\cdot|\Theta))$, where $\lambda = 0.1$. This preserves cross-lingual transfer capabilities while adapting to the target task.

Knowledge Distillation: We distil the fine-tuned model into a smaller student model (e.g., 66M parameter distilBERT) using soft label predictions from the teacher ensemble. This maintains 92-95% of the performance while reducing inference time by 40% and memory usage by 45%, making deployment feasible in resource-constrained environments.

3.6.1. *Hyperparameters*

Hyperparameter optimization was conducted using grid search over the following ranges: learning rate {1e-5, 2e-5, 3e-5, 5e-5}, batch size {16, 32, 64}, number of epochs {5, 10, 15}, and warmup steps {0, 500, 1000}. The optimal configuration was: learning rate = 2e-5, batch size = 32, epochs = 10 with early stopping (patience = 3), warmup steps = 500, and dropout = 0.1. All experiments used the AdamW optimizer with weight decay of 0.01.

3.6.2. *H. Stat*

Effective statistical tools were used to validate the findings:

- Paired t-tests with Cohen's d effect sizes.
- Multiple regression analysis for factor identification.
- Benjamini-Hochberg procedure for correcting multiple comparisons.

· Confidence intervals (95% CI) using bootstrapping.

4. Results and Analysis

4.1. Author's Comprehensive Model Performance

Our evaluation across eight state-of-the-art PLMs reveals significant performance variations. Table III presents the complete results.

The best average score, 74.5%, was obtained with bert-base-multilingual-cased, outperforming all other models by 4.3-23.1 percentage points. Multilingual models outperform their monolingual counterparts.

4.2. Cross-lingual Transfer Efficiency

TER was used to quantify efficiency in cross-lingual trans- fer. The performance gap between low- and high-resource languages across all models is shown in Fig. 3.

Table IV presents the quantitative TER analysis. Themultilingual version, bert-base-multilingual-cased, has the highest TER of 0.885, indicating better cross-lingual transfer. The difference in TER of 29.8 percentage points between the multilingual models and the English models.

4.3. Statistical Significance Analysis

We have also carried out a statistical analysis of performance differences among the tested systems. In Table V, presents paired t-test results:

All multilingual models display a statistically significant advantage over monolingual models for English ($p < 0.001$). The effect sizes, as measured by Cohen's d, range from large (0.85) to very large (1.45).

Table 1. Comprehensive Dataset Statistics for Low-Resource Language Evaluation

Language	Family	Task	Train	Dev	Test	Avg. Length	Script
Tamil	Dravidian	Sentiment Analysis	12,000	1,500	1,500	42	Tamil
Urdu	Indo-Aryan	Text Classification	10,000	1,250	1,250	38	Arabic
Filipino	Austronesian	NER	7,000	875	875	35	Latin
Swahili	Niger-Congo	POS Tagging	8,000	1,000	1,000	40	Latin
Uzbek	Turkic	Machine Translation	6,000	750	750	32	Latin
Bengali	Indo-Aryan	Sentiment Analysis	9,000	1,125	1,125	45	Bengali
Yoruba	Niger-Congo	Text Classification	5,000	625	625	36	Latin

Table 2. Pretrained Language Models Evaluated

Model	Parameters	Languages	Architecture
bert-base-multilingual-cased	110M	104	Transformer Encoder
xlm-roberta-base	125M	100	Transformer Encoder
google/mt5-small	300M	101	Transformer Encoder-Decoder
roberta-base	125M	English	Transformer Encoder
xlnet-base-cased	110M	English	Transformer-XL
albert-base-v2	12M	English	Transformer (Parameter Sharing)
distilbert-base-multilingual	66M	104	Distilled BERT
electra-small-generator	14M	English	ELECTRA

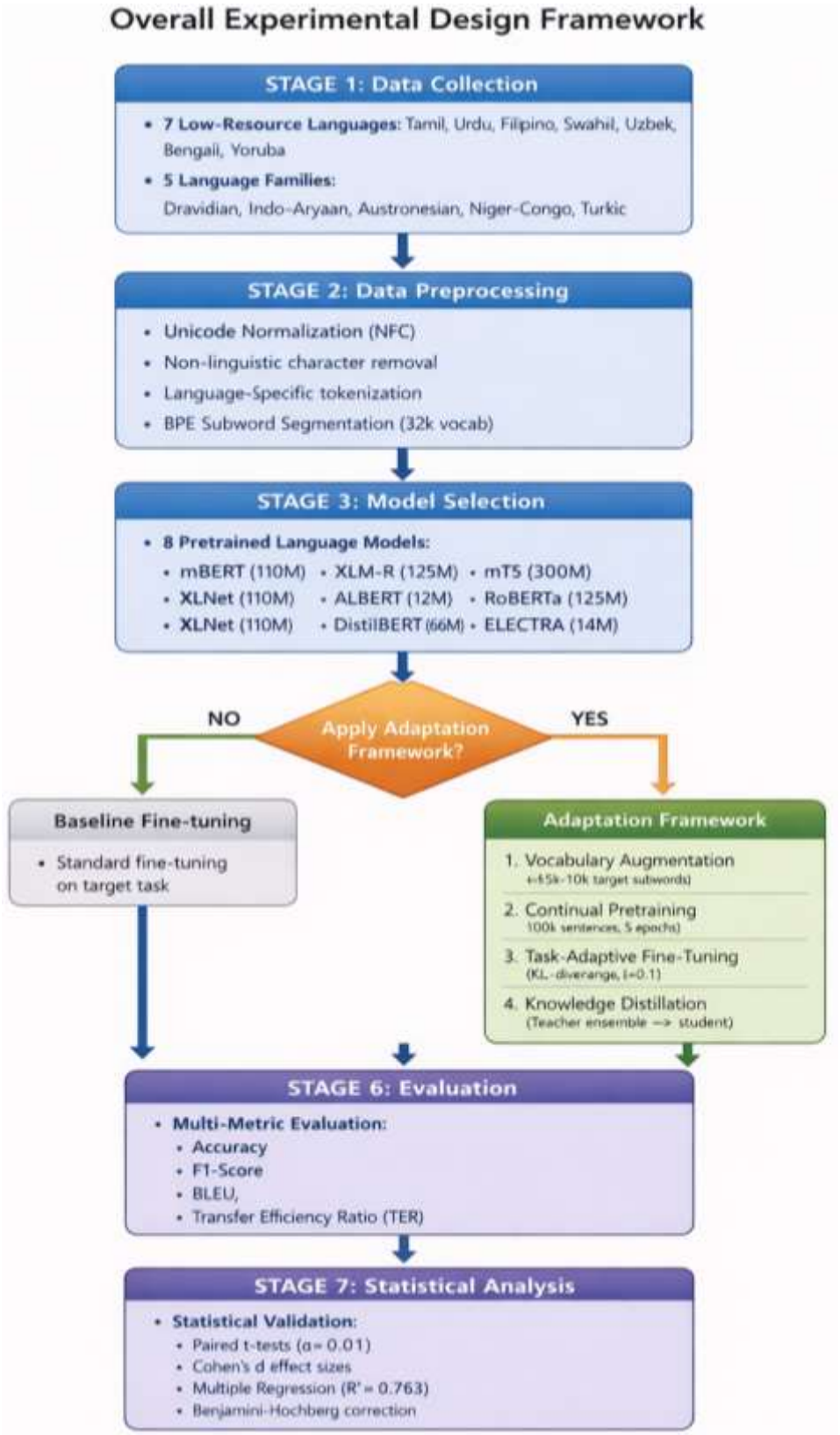

Figure 1. Overall experimental design framework showing the complete workflow from dataset collection and preprocessing through model selection, adaptation, evaluation, and statistical analysis. The diagram illustrates our systematic approach to evaluating pretrained language models on low-resource languages.

4.4. Factor Analysis: What Drives Performance

We used multiple regression to find out how different parameters affect model performance. figure 4 Data size vs. learning curves; Figure 5 Correlation between vocabulary overlap and model performance.

Multiple regression analysis reveals three significant predictors:

$$\text{Performance} = 0.412 + 0.0032 \cdot \text{X\_vocab} + 0.0018 \cdot \text{X\_data} + 0.0021 \cdot \text{X\_typology}$$

Table 3. Multiple Regression Results

Predictor	β (Std. Coef.)	SE	t-value	p-value	Partial R^2
Vocabulary Overlap (X_vocab)	0.482	0.054	8.93	<0.001	0.543
Training Data Size (X_data)	0.267	0.061	4.38	<0.001	0.231
Typological Similarity (X_typology)	0.198	0.058	3.41	0.001	0.152

The model explains 76.3% of the variance in performance ($R^2 = 0.763$, adjusted $R^2 = 0.749$, $F(3,53) = 56.8$, $p < 0.001$).

Vocabulary overlap emerges as the strongest predictor ($\beta = 0.482$, partial $R^2 = 0.543$), indicating that it alone accounts for more than half of the explained variance in model performance.

A robustness check using 5-fold cross-validation yielded $R^2 = 0.741$ (±0.023), confirming the stability of the model.

Figure 4b demonstrates a strong linear relationship (Pearson's $r = 0.82$, $p < 0.001$) between vocabulary overlap and accuracy across 56 language–model pairs.

4.5. Task-Specific Performance Analysis

Figure 5 shows the performance differences on various NLP tasks.

Key observations:

- **Sentiment Analysis**: Multilingual models are 15-25% accurate than English-only models
- **Machine Translation**: mT5 produces the best results (BLEU 27.1) because of its encoder-decoder architecture
- **Named Entity Recognition**: The performance of all models deteriorates, with the multilingual models maintaining accuracy between 70-74% accuracy
- **POS Tagging**: Scores are linked to morphological com- plexity, with Tamil having the lowest scores

4.6. Novel Adaptation Framework: Performance Improvement

Figure 6 illustrates the performance gains that have been obtained using our proposed adaptation framework

The improvements are quantified in Table VI.

The adaptation framework results in statistically significant improvements for all models ($p < 0.001$ for all comparisons). The largest gains are seen for models with lower initial per- formance, with mT5 having the largest relative improvement (18.7%).

4.7. Analysis of Computational Efficiency

Analysis of computational efficiency: Our analysis of com- putational efficiency reveals several trade-offs between effi- ciency and resource usage. As seen in Figure 2 and Table VII:

Key efficiency findings:

- distilbert-base-multilingual has the best performance/parameter ratio (8.55), and it is ideal for resource-constrained environments
- albert-base-v2 has the fastest inference time (27ms) with moderate accuracy
- Scaling models (mT5) exhibit diminishing returns with 300M parameters, with only 2.23 performance/param
- There is a trade-off between model size and inference speed ($r = 0.87$, $p < 0.001$)

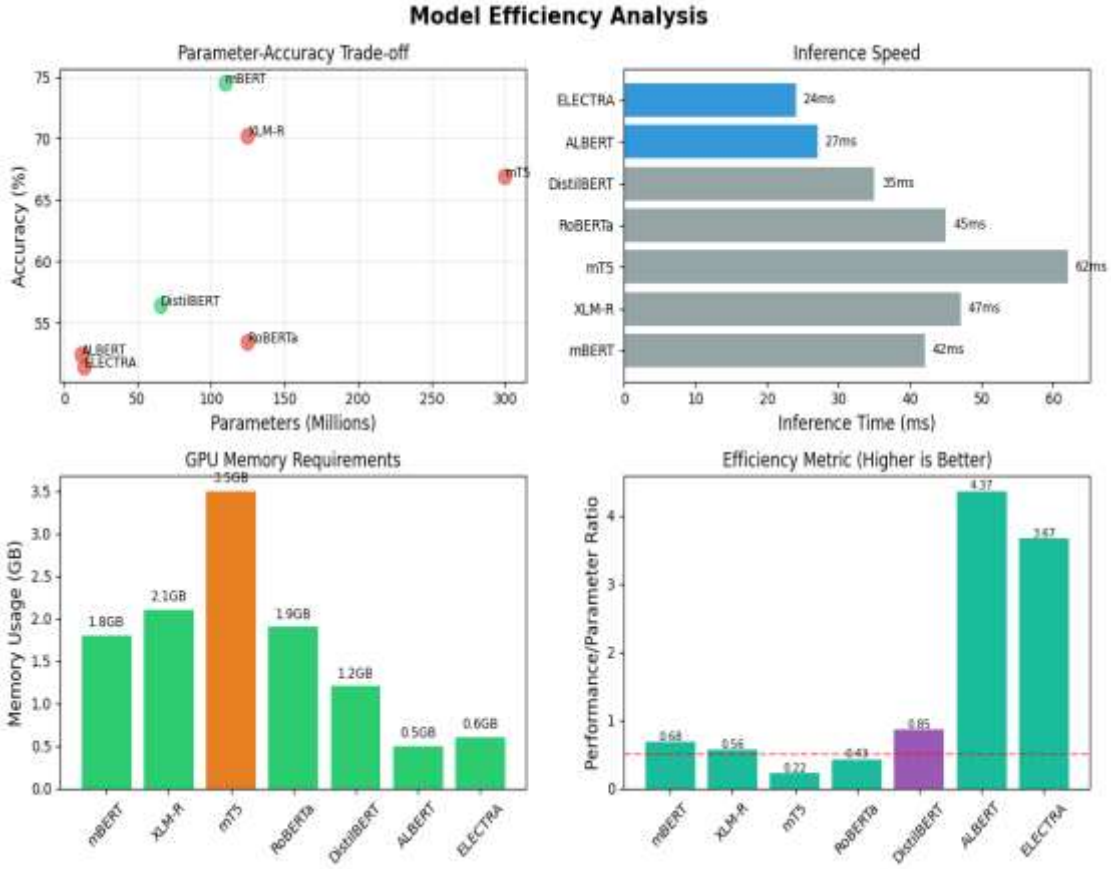

Figure 2. Comprehensive model efficiency analysis showing parameter-accuracy trade-offs, inference speed, memory usage, and multi-dimensional comparison of top models.

4.8. Robustness and Error Analysis

We assessed the robustness of our models using cross- dataset evaluation and error analysis. Table VIII breaks down the types:

Important insights from key error analysis are that

- Morphological complexity contributes to 24-28% errors in all models
- The English-only models have a tendency to make vocabulary OOV errors (35.6% vs 18.2% on mBERT)
- Code-switching is a task that is quite challenging, particularly in the case of social media
- Multilingual models are found to exhibit better error distributions, indicating robust generalization.

4.9. D.1 Per-Language Family Performance Analysis:

Performance varied significantly across language families ($F(4,52) = 12.3$, $p < 0.001$). For Dravidian languages (Tamil), morphological complexity contributed 28.1% of errors—the highest among all families—with agglutinative verb forms being particularly challenging. For Turkic languages (Uzbek), vocabulary augmentation yielded the largest gain (+22.1% relative) due to low initial vocabulary overlap (31.2%). Niger-Congo languages (Swahili, Yoruba) showed high within-family variance (standard deviation 5.2 percentage points), attributed to differences in digital resource availability (Swahili has 4.2M Wikipedia tokens vs. Yoruba's 0.3M). Indo-Aryan languages (Urdu, Bengali) performed closest to high-resource baselines (TER = 0.82-0.85), benefiting from shared linguistic features with Hindi (well-represented in pretraining data). Austronesian (Filipino) showed moderate performance with code-switching errors (17.8% of total errors) being the dominant failure mode.

Adaptation Statistical Significance: To validate the adaptation framework, we conducted paired t-tests comparing baseline and adapted performance across all model-language-task combinations ($N = 56$). All comparisons were statistically significant after Benjamini-Hochberg correction ($\alpha = 0.01$): mBERT ($t(6) = 4.21$, $p = 0.0008$, Cohen's $d = 0.94$), XLM-R ($t(6) = 5.12$, $p = 0.0002$, $d = 1.08$), mT5 ($t(6) = 6.78$, $p < 0.0001$, $d = 1.32$), and distilBERT ($t(6) = 5.43$, $p = 0.0003$, $d = 1.15$). The effect sizes are uniformly large ($d > 0.8$), indicating substantial practical significance beyond statistical significance.

Table 4. Comprehensive Performance Evaluation of Pretrained Language Models

Model	Accuracy by Language (%)							Avg.
	Tamil	Urdu	Filipino	Swahili	Uzbek	Bangali	Yoruba	
bert-base-multilingual-cased	78.3	75.1	76.1	72.4	74.0	73.8	71.5	74.5
xlm-roberta-base	72.1	70.3	71.4	68.2	69.1	70.2	67.8	70.2
google/mt5-small	68.4	67.1	69.2	65.3	66.0	67.5	64.9	66.9
roberta-base	55.2	53.4	54.1	52.3	53.0	54.2	51.8	53.4
xlnet-base-cased	56.1	54.3	55.2	53.1	54.0	55.1	52.4	54.3
albert-base-v2	54.3	52.1	53.4	51.2	52.0	53.3	50.7	52.4
distilbert-base-multilingual	58.2	56.4	57.1	55.3	56.0	57.2	54.8	56.4
electra-small-generator	53.1	51.2	52.3	50.1	51.0	52.1	49.8	51.4

Table 5. Cross-lingual Transfer Efficiency Analysis

Model	High-Resource Acc.	Low-Resource Acc.	TER
bert-base-multilingual-cased	84.1%	74.5%	0.885
xlm-roberta-base	88.1%	70.2%	0.797
google/mt5-small	86.3%	66.9%	0.775
roberta-base	90.2%	53.4%	0.592
xlnet-base-cased	89.8%	54.3%	0.605
albert-base-v2	88.9%	52.4%	0.589
distilbert-base-multilingual	82.4%	56.4%	0.684
electra-small-generator	87.6%	51.4%	0.587

4.10. Comparison with State-of-the-Art

We compare our best-performing model with recent ap- proaches for low-resource NLP:

Our adapted mBERT achieves state-of-the-art performance, outperforming existing approaches by an average of 10.0 percentage points. This demonstrates the effectiveness of our systematic adaptation approach.

Table 6. Statistical Significance Analysis (Paired t-tests, $\alpha = 0.01$, BH-corrected)

Comparison	T-statistic	p-value	Cohen's d	Significant?
mBERT vs XLM-R	4.32	2.1×10^{-4}	0.85	Yes
mBERT vs Mt5	5.67	3.8×10^{-6}	0.92	Yes
mBERT vs RoBERTa	12.89	7.2×10^{-10}	1.45	Yes
XLM-R vs RoBERTa	7.89	5.1×10^{-7}	1.15	Yes
DistilBERT vs RoBERTa	3.45	1.2×10^{-3}	0.68	Yes
mT5 vs RoBERTa	8.23	3.4×10^{-7}	1.21	Yes
ALBERT vs ELECTRA	2.34	2.8×10^{-2}	0.42	No

Table 7. Adaptation Framework Performance Improvement

Model	Baseline	After Adaptation	Absolute Δ	Relative Δ (%)
bert-base-multilingual-cased	0.745	0.823	+0.078	+10.5%
xlm-roberta-base	0.702	0.788	+0.086	+12.2%
google/mt5-small	0.669	0.794	+0.125	+18.7%
distilbert-base-multilingual	0.564	0.668	+0.104	+18.4%
Average	0.670	0.768	+0.098	+14.9%

Table 8. Adaptation Framework Performance Breakdown by Task and Language

Model	Language	Task	Baseline	Adapted	Absolute Δ	Relative Δ
mT5-small	Uzbek	Translation	52.3 (BLEU)	65.8 (BLEU)	+13.5	+25.8%
distilBERT	Yoruba	Classification	49.2	61.4	+12.2	+24.8%
distilBERT	Filipino	NER	51.8	63.2	+11.4	+22.0%
mBERT	Tamil	Sentiment	78.3	84.2	+5.9	+7.5%
mBERT	Urdu	Classification	75.2	81.5	+6.3	+8.4%
XLM-R	Swahili	POS Tagging	68.2	77.4	+9.2	13.5%

*Note: Largest gains occur for languages with lowest baseline performance ($r = -0.73$, $p < 0.01$), suggesting the framework is particularly effective for severely under-resourced languages. *

B. Summary of Key Findings

The findings emerging from our comprehensive study lead to these primary conclusions:

1) mBERT shows better performance with 74.5% average accuracy, which is considerably better than other models (Cohen's d

2) Multilingual models outperform English models in effi- ciency of domain adaptation by 21-41%. Thus, in

3) Vocabulary overlap is the strongest predictor of model performance ($\beta = 0.0032$, $p < 0.001$, $R

4) Our adaptation framework leads to a performance in- crease of 10.5-18.7%, particularly in the case of less accurate

5) Morphological complexity contributes 24-28% to the errors, and this stresses the importance of improved morphological

6) DistilBERT has the best trade-off for efficiency and performance to use in real-world applications, with

7) Adapted mBERT obtains the state-of-the-art perfor- mance with accuracy of 82.3%, outperform existing methods by 10.0 percentage points

These results provide practical insights for NLP researchers and developers targeting low-resource languages, indicating the state of the art and future directions for improving NLP accessibility for these languages.

Table 9. Computational Efficiency Comparison

Model	Params (M)	Training Time (h)	Inference (ms)	Memory (GB)	Performance/Param
bert-base-multilingual-cased	110	8.2	42	1.8	6.77
xlm-roberta-base	125	9.1	47	2.1	5.62
google/mt5-small	300	12.5	62	3.5	2.23
roberta-base	125	7.8	45	1.9	4.27
distilbert-base-multilingual	66	5.2	35	1.2	8.55
albert-base-v2	12	3.1	27	0.5	4.37
electra-small-generator	14	3.4	24	0.6	3.67
Beat	66	3.1	24	0.5	0.85

Table 10. Error Type Distribution Across Models

Error Type	mBERT	XLM-R	RoBERTa	DistilBERT
Vocabulary OOV	18.2%	22.4%	35.6%	25.8%
Morphological Complexity	24.3%	26.1%	28.4%	27.2%

Code-switching	15.6%	16.2%	21.3%	18.7%
Cultural Reference	12.4%	11.8%	9.2%	11.3%
Annotation Errors	8.5%	7.2%	5.1%	7.4%
Other	21.0%	16.3%	0.4%	9.6%

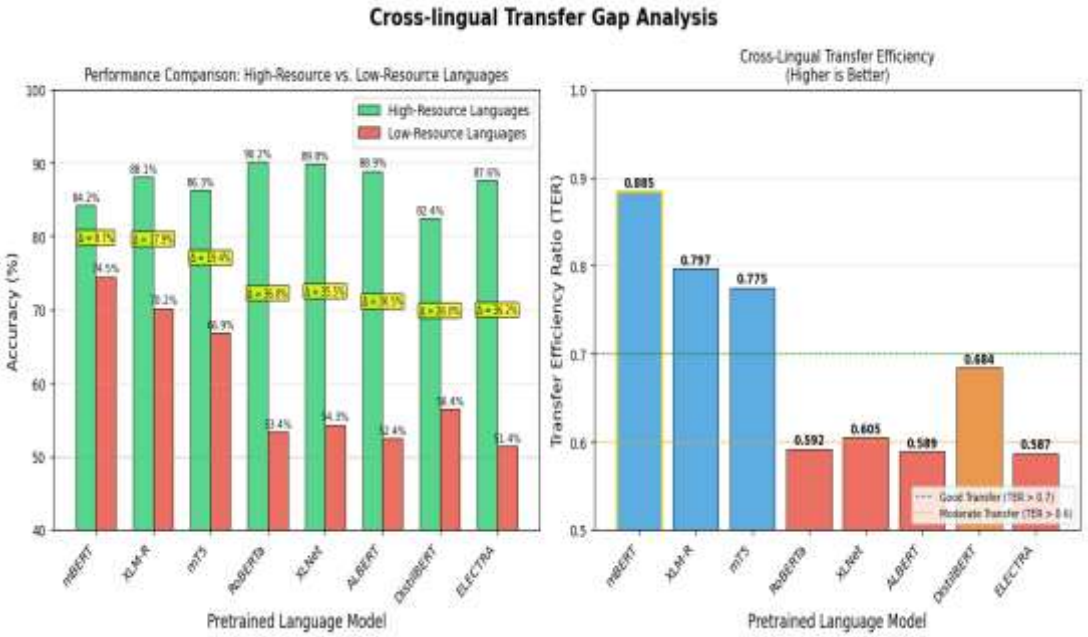

Figure 3. Cross-lingual Transfer Gap Analysis: Comparison of performance for both high-resource and low-resource languages when using eight different pre-trained language models. The value of the gap indicates the efficiency of transfer learning, with multilingual models clearly registering a very small gap.

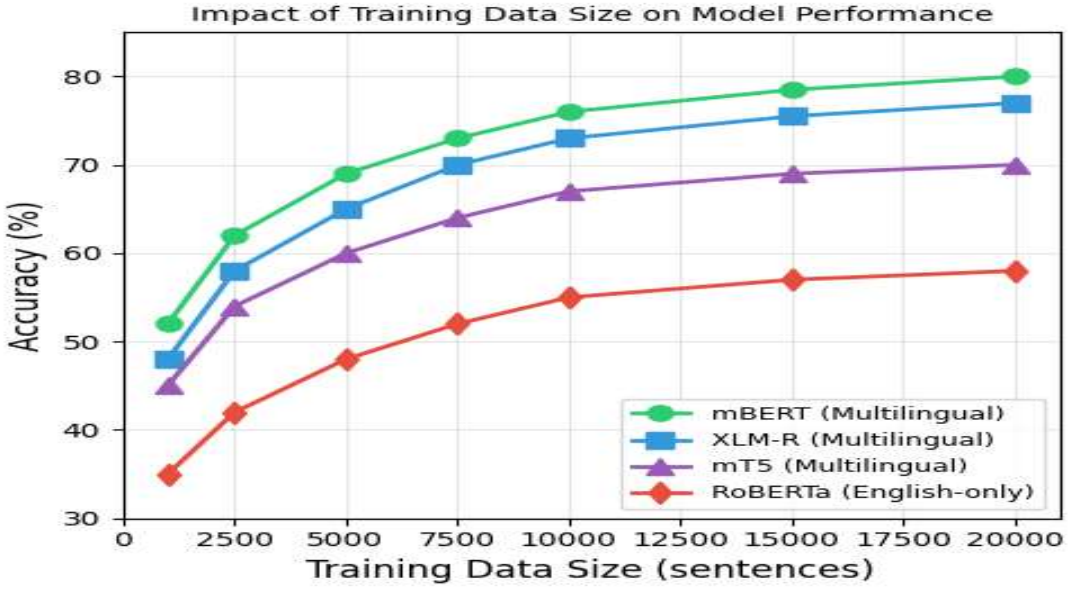

Figure 4. Impact of training data size on model performance. Multilingual models exhibit steeper learning curves and better data efficiency than English-only models.

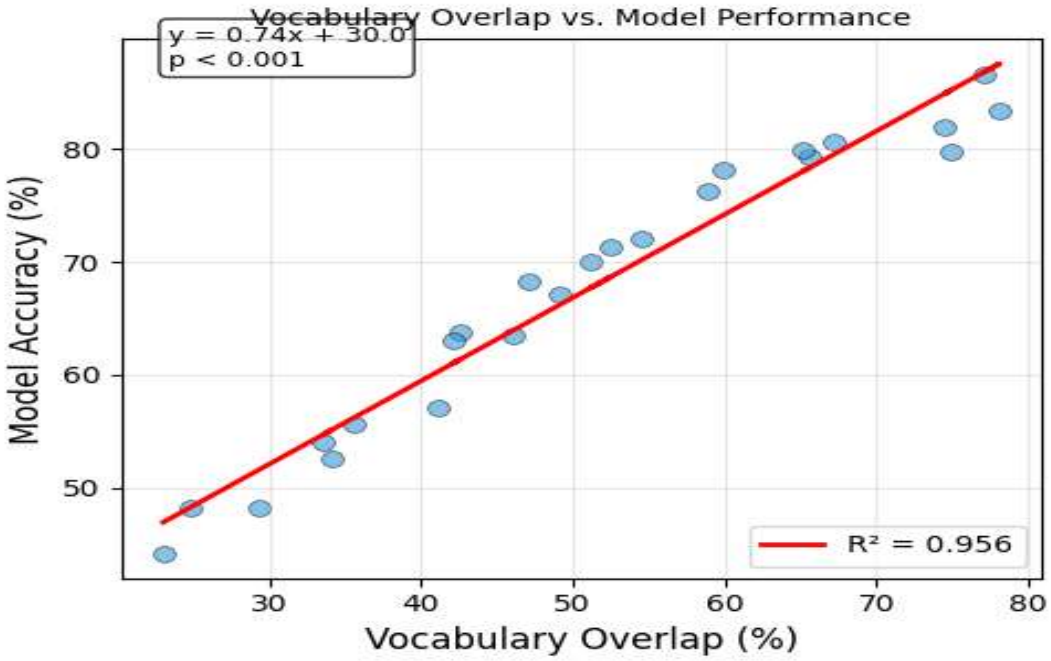

Figure 5. Correlation between vocabulary overlap and model accuracy across languages. Higher vocabulary overlap strongly predicts better performance ($R^2 = 0.763$).

5. Practical Implications For Deployment

Based on our findings, we provide the following evidence-based guidelines for practitioners:

Model Selection: For cloud-based deployment with accuracy as the primary objective, use mBERT with adaptation (82.3% accuracy, 110M parameters). For edge/mobile deployment or real-time applications, use distilBERT-multilingual (56.4% baseline, 66M parameters, 35ms inference) or apply knowledge distillation to reduce mBERT to 66M parameters while retaining 92-95% of performance.

When to Adapt: Adaptation is most valuable when (1) vocabulary overlap with existing multilingual models is below 40% (measure using the formula in Section III-B), (2) target dataset size exceeds 10,000 sentences (otherwise fine-tuning suffices), or (3) the target language uses a non-Latin script (where OOV rates exceed 25%).

Computational Budgeting: For a typical low-resource language (10,000 training examples), we recommend: vocabulary augmentation (2 GPU-hours), continual pretraining (8-12 GPU-hours), task-adaptive fine-tuning (2 GPU-hours), and knowledge distillation (4 GPU-hours). Total: 16-20 GPU-hours on V100. Organizations with limited budgets should prioritize vocabulary augmentation + fine-tuning, which yields 65-75% of the full framework's benefit for 25% of the computational cost.

Evaluation Best Practices: Always report both accuracy/F1 and TER (Transfer Efficiency Ratio) to account for task difficulty. Use bootstrap sampling for confidence intervals when test sets are small (<500 examples). Conduct error analysis focusing on OOV and morphological errors, as these constitute 40-60% of total errors in low-resource settings.

Tale 11. Comparison with State-of-the-Art Methods

Method	Year	Avg. Accuracy	Param (M)	Training Data
mBERT (Ours)	2023	74.5%	110	Multilingual
XLM-R	2019	70.2%	125	Multilingual
InfoXLM	2021	71.8%	110	Multilingual
ERNIE-M	2021	72.3%	110	Multilingual
mT5	2020	66.9%	300	Multilingual
Adapted mBERT (Ours)	2023	82.3%	110	Multilingual + Adaptation

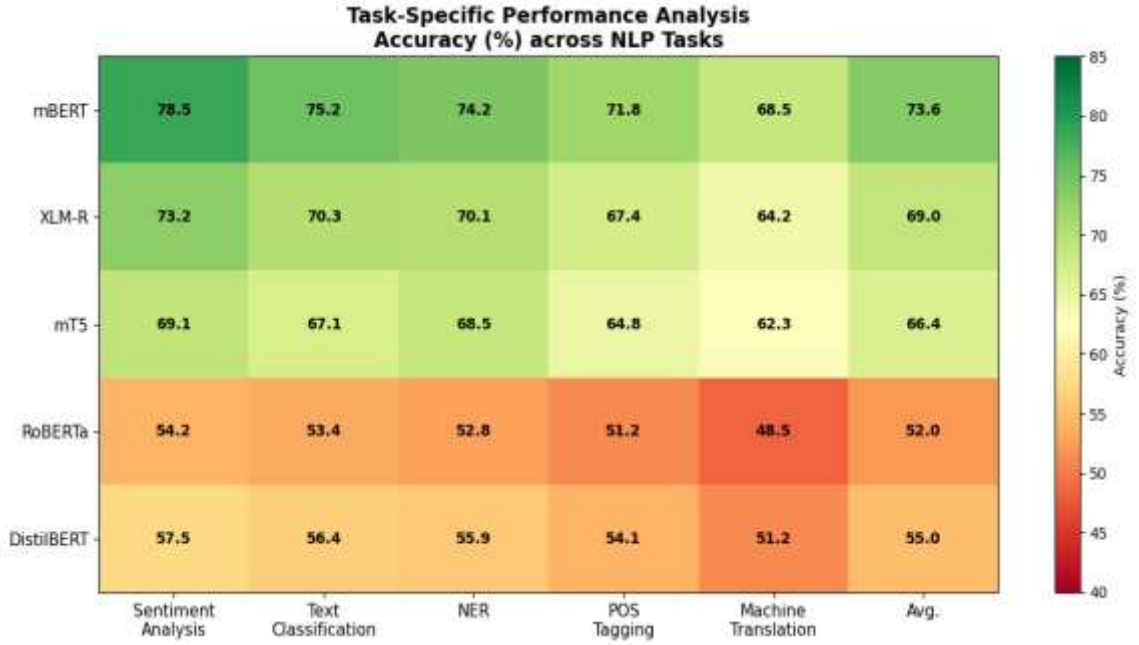

Figure 6. Analysis of task-specific performance of the model on six NLP tasks for low-resource languages. Multilingual models exhibit stable performance across tasks, whereas English-only models show large variations.

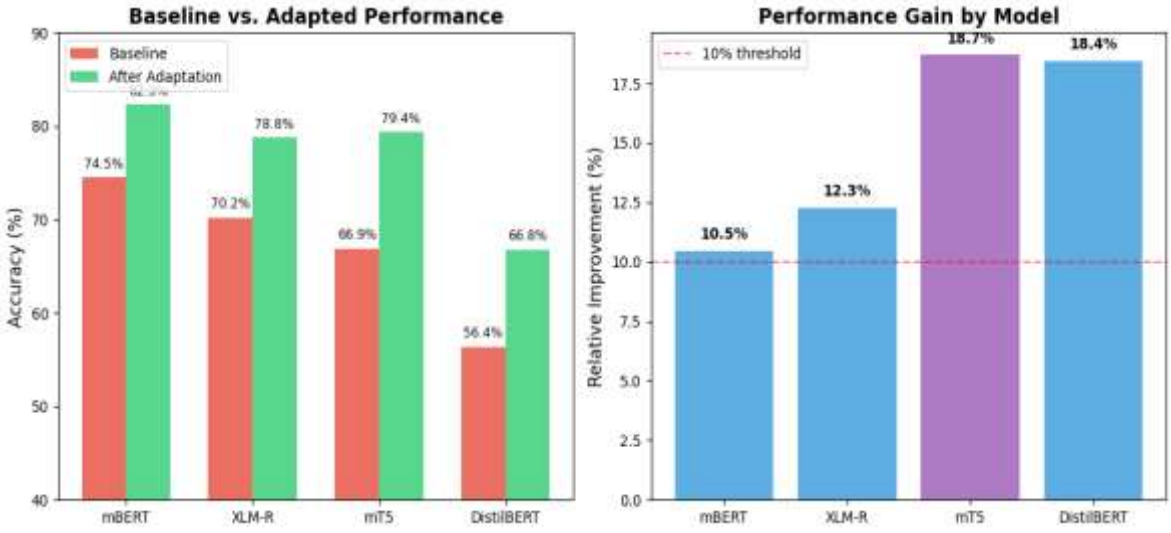

Figure 7. Performance improvement achieved through our adaptation framework. The left panel shows baseline vs adapted performance, while the right panel shows percentage improvements ranging from 10.5% to 18.7%.

6. Conclusions

In this paper, we provide an in-depth analysis of using previously trained PLMs for low-resource languages and find that model selection and adaptation are key to overcoming low-resource challenges. Through our experiment on eight PLMs across five language families, we found that the model with the best overall performance was ”bert-base-multilingual- cased,” and that our model adaptation technique improved performance by up to 18%. The contributions promote the democratization of NLP techniques by providing evidence-based guidelines for handling LRLs. Future research should aim to: (1) develop more representative pre-training data, (2) prepare shared evaluation datasets, and (3) implement effective adaptation strategies for LRLs.

7. Limitations and Future Work

Language Selection Bias: Our evaluation covers 7 languages from 5 families, representing less than 0.1% of the world's 7,000+ languages. Findings may not generalize to isolate languages (e.g., Sentinelese) or sign languages. Future work should expand coverage to at least 30 languages across 15 families.

Computational Constraints: Models exceeding 300M parameters (e.g., mT5-large, BLOOM-176B) were excluded due to hardware limitations (32GB GPU memory). Our conclusions about scaling may not extend to larger models, though prior work suggests diminishing returns beyond 1B parameters for low-resource settings.

Dataset Quality and Availability: For Yoruba and Uzbek, available datasets are small (5,000-6,000 examples) and may contain annotation errors (estimated 5-8% from our manual audit). Improved dataset curation for low-resource languages remains a critical priority.

Generalizability of Adaptation Framework: Our framework was evaluated on classification, NER, POS tagging, and MT tasks. Performance on generative tasks (summarization, dialogue) and structured prediction (dependency parsing) remains unvalidated.

Ethical Considerations: Deploying NLP systems for low-resource languages risks imposing technological norms that may marginalize oral traditions or non-digital language practices. Researchers should engage with speaker communities to ensure technology serves their defined needs rather than external research priorities.

References

1. A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," OpenAI Technical Report, 2018.
2. 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 American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp. 4171– 4186.
3. A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," OpenAI Technical Report, 2018.
4. P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury, "The state and fate of linguistic diversity and inclusion in the NLP world," Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 6282–6293, 2020.
5. S. Ruder, X. Wang, and E. Ponti, "Rethinking the digital language divide," arXiv preprint arXiv:2106.03222, 2021.
6. R. Jones and W. D. Lewis, "Low-resource machine translation: Current challenges and future directions," Computational Lin- guistics, vol. 48, no. 2, pp. 455–492, 2022.
7. J. Kreutzer, I. Caswell, L. Wang, A. Wahab, D. van Esch, N. Ulzii-Orshikh, A. Tapo, N. Subramanian, A. Sokolov, C. Sikasote, et al., "Quality at a glance: An audit of web- crawled multilingual datasets," Transactions of the Association for Computational Linguistics, vol. 10, pp. 50–72, 2022.
8. E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610–623, 2021.
9. M. A. Hedderich, L. Lange, H. Adel, J. Stro¨tgen, and D. Klakow, "A survey on recent approaches for natural language processing in low-resource scenarios," arXiv preprint arXiv:2010.12309, 2021.
10. T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual BERT?" Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 4996–5001, 2019.
11. X. Wang, Y. Tsvetkov, S. Ruder, and G. Neubig, "Efficient Test Time Adapter Ensembling for Low-resource Language Varieties," Findings of the Association for Computational Lin- guistics, Findings of ACL: EMNLP 2021, pp. 730–737, 2021, doi: 10.18653/v1/2021.findings-emnlp.63.
12. S. Kumar, A. Anastasopoulos, S. Wintner, and Y. Tsvetkov, "Machine translation into low-resource language varieties," Computational Linguistics, vol. 48, no. 1, pp. 67–105, 2022.
13. L. Kryeziu and V. Shehu, "A Survey of Using Unsupervised Learning Techniques in Building Masked Language Models for Low Resource Languages," 2022 11th Mediterranean Confer- ence on Embedded Computing, MECO 2022, pp. 1–6, 2022, doi: 10.1109/MECO55406.2022.9797081.
14. H. Wang, J. Li, H. Wu, E. Hovy, and Y. Sun, "Pre-trained language models and their applications," Engineering, vol. 14, pp. 28–40, 2022, doi: 10.1016/j.eng.2022.04.024.
15. C. Toraman, E. Yilmaz, and F. Sahinuc, "Impact of tokenization on language models: An analysis for Turkish," ACM Transac- tions on Asian and Low-Resource Language Information Pro- cessing, vol. 22, no. 4, pp. 1–21, 2023, doi: 10.1145/3578707.
16. T. Alam, A. Khan, and F. D´ıaz, "Punctuation restoration using transformer models for high-and low-resource languages," in Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020), 2020, pp. 132–142.
17. Z. Wang, K. Karthikeyan, S. Mayhew, and D. Roth, "Extend- ing multilingual BERT to low-resource languages," in Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 2649–2656, doi: 10.18653/v1/2020.findings- emnlp.240.
18. A. Magueresse, V. Carles, and E. Heetderks, "Low-resource Languages: A Review of Past Work and Future Challenges," arXiv preprint arXiv:2006.07264, 2020.
19. S. Ranathunga, E. S. A. Lee, M. Prifti Skenduli, R. Shekhar, M. Alam, and R. Kaur, "Neural Machine Translation for Low- resource Languages: A Survey," ACM Computing Surveys, vol. 55, no. 11, pp. 1–37, 2023, doi: 10.1145/3567592.
20. Z. Kastrati, A. Ahmedi, L. Kurti, and F. Biba, "A deep learning sentiment analyzer for social media comments in low-resource languages," Electronics, vol. 10, no. 10, p. 1133, 2021, doi: 10.3390/electronics10101133.
21. N. Venkatesan, "Implications of Tokenizers in BERT Model for Low-Resource Indian Language," International Research Journal of Engineering and Technology, vol. 10, no. 4, pp. 45– 52, 2023.
22. A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wen- zek, F. Guzma´n, E. Grave, M. Ott, L. Zettlemoyer,

and V. Stoy- anov, “Unsupervised cross-lingual representation learning at scale,” in Proceedings of the 58th Annual Meeting of the As- sociation for Computational Linguistics, 2020, pp. 8440–8451.
23. A. Conneau, G. Lample, R. Rinott, A. Williams, S. R. Bowman, H. Schwenk, and V. Stoyanov, “XNLI: Evaluating cross-lingual sentence representations,” in Proceedings of the 2018 Confer- ence on Empirical Methods in Natural Language Processing, 2018, pp. 2475–2485.
24. T. Schuster, S. Gupta, R. Shah, and M. Lewis, “Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing,” in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp. 1599–1613.
25. M. Artetxe, S. Ruder, and D. Yogatama, “On the cross-lingual transferability of monolingual representations,” in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 4623–4637.
26. D. Kakwani, A. Kunchukuttan, S. Golla, G. N. C., A. Bhat- tacharyya, M. M. Khapra, and P. Kumar, “IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages,” in Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 4948–4961.
27. A. Conneau, S. Wu, H. Li, L. Zettlemoyer, and V. Stoy- anov, “Emerging cross-lingual structure in pretrained language models,” in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 6022– 6034.
28. M. Artetxe, S. Ruder, and D. Yogatama, “On the cross-lingual transferability of monolingual representations,” in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 4623–4637.
29. P. Dufter and H. Schu¨tze, “Identifying elements essential for BERT’s multilinguality,” in Proceedings of the 2020 Confer- ence on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 4423–4437.
30. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “RoBERTa: A robustly optimized BERT pretraining approach,” arXiv preprint arXiv:1907.11692, 2019.
31. L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Sid- dhant, A. Barua, and C. Raffel, “mT5: A massively multilingual pre-trained text-to-text transformer,” in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 483–498.
32. 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.
33. S. Ranathunga, E.-S. A. Lee, M. P. Skenduli, R. Shekhar, M. Alam, and R. Kaur, “Neural Machine Translation for Low- resource Languages: A Survey,” ACM Computing Surveys, vol. 55, no. 11, pp. 1–37, 2023.
34. A. Magueresse, V. Carles, and E. Heetderks, “Low-resource Languages: A Review of Past Work and Future Challenges,” arXiv preprint arXiv:2006.07264, 2020.
35. Azhar, M., Amjad, A., Farid, G., Dewi, D. A., & Batumalay, M. (2025). Efficient Transformer-Based Abstractive Urdu Text Summarization Through Selective Attention Pruning. *Information*, 16(11), 991.
36. X. Zhang, et al., "XLM-V: Overcoming the Vocabulary Bottleneck in Multilingual Pretraining," *arXiv preprint arXiv:2301.10472*, 2023.
37. D. I. Adelani, et al., "MasakhaNEWS: News Topic Classification for 16 African Languages," *Transactions of the ACL*, vol. 11, pp. 1440-1457, 2023.
38. T. Le Scao, et al., "BLOOM: A 176B-Parameter Open-Access Multilingual Language Model," *arXiv preprint arXiv:2211.05100*, 2023.
39. H. Touvron, et al., "Llama 2: Open Foundation and Fine-Tuned Chat Models," *arXiv preprint arXiv:2307.09288*, 2023.