

A Performance-efficiency Analysis of Transformer Models for Code-mixed Hausa Sentiment Data

Muhammed Bashir Muazu

`mbmuazu@abu.edu.ng`

Ahmadu Bello University <https://orcid.org/0000-0003-2003-4761>

Abdulkadir Bayero

`abdulkadirbayero6@gmail.com`

Ahmadu Bello University <https://orcid.org/0009-0001-5849-5574>

Abdullahi Ahmad

Ahmadu Bello University

Sufiyanu Ibrahim

Ahmadu Bello University

Dahir Ibrahim Dahir

Ahmadu Bello University

Research Article

Keywords: Hausa NLP, Sentiment Analysis, Transformer Models, Efficiency, Low-Resource NLP, Code-mixed Text, Deployment

Posted Date: March 2nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8982792/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

A Performance-Efficiency Analysis of Transformer Models for Code-Mixed Hausa Sentiment Data

Muhammed Bashir Muazu, Abdulkadir Bayero, Abdullahi Ahmad, Sufiyanu Ibrahim, and Dahir Ibrahim Dahir

Department of Computer Engineering,
Ahmadu Bello University, Zaria

Abstract. The deployment of large language models (LLMs) in African contexts offers significant opportunities for societal innovation, but is often hindered by their substantial computational requirements. This challenge is particularly acute for code-mixed Hausa sentiment analysis, where the state-of-the-art (SOTA) model achieves high accuracy at a significant cost in model size and inference speed. This paper addresses this performance-efficiency trade-off by presenting the first comprehensive benchmark of the SOTA model against a suite of smaller, more efficient alternatives, including specialized African-centric and generalist models. Using the established NaijaSenti benchmark, our experiments reveal that a compact, specialized model (`castorini/afriberta_small`) retains over 82% of the SOTA’s F1-score (0.761 vs. 0.928) while being 2.25 times faster at inference. Furthermore, our results demonstrate that specialized pre-training is a critical factor, as the small African-centric model significantly outperforms its generalist counterpart. We conclude that `castorini/afriberta_small` represents the optimal "sweet spot" for practical deployment, and we recommend it for applications where a balance between high performance and computational efficiency is required. Our findings provide a data-driven guide for practitioners building scalable NLP solutions in the African context.

Keywords: Hausa NLP, Sentiment Analysis, Transformer Models, Efficiency, Low-Resource NLP, Code-mixed Text, Deployment

1 Introduction

Recent advances in Large Language Models (LLMs) present a significant opportunity to foster societal innovation and address long-standing challenges in digital inclusion. In the African context, these models have the potential to bridge linguistic divides, power new applications in sectors such as healthcare and agriculture, and provide access to information to millions of people [1]. However, the significant computational resources required to train and deploy these models present a substantial barrier, particularly in resource-constrained environments. This challenge of real-world AI deployment is a critical hurdle that

must be addressed to realize the full potential of this technology, particularly in resource-constrained African contexts where computational infrastructure may be limited [18].

A key application for this technology is sentiment analysis, particularly for understanding public discourse on social media. Hausa, a language spoken by over 80 million people, is a prime example where this is both highly valuable and technically challenging. Social media text in this context is often "code-mixed," blending Hausa and English within a single sentence, which poses a significant challenge for traditional NLP models [2]. Effectively analyzing sentiment in this setting requires models with a deep contextual understanding of both languages.

Significant progress in this area was made by the authors of the work that established" or "the work by [3] that developed. Their work established a powerful state-of-the-art (SOTA) baseline, demonstrating that a large transformer model pre-trained on African languages (AfriBERTa) could achieve over 92% accuracy on this task. However, the authors themselves noted the practical limitations of this approach, highlighting that large transformer models are computationally expensive. This trade-off between high performance and the high cost of deployment creates a critical gap for practitioners seeking to build scalable, real-world applications.

This paper directly addresses the performance-efficiency trade-off for Hausa sentiment analysis. We present the first comprehensive benchmark of smaller, more efficient transformer models against the established SOTA. Our goal is to identify a "sweet spot" model that provides a practical balance for real-world deployment. Specifically, our contributions are threefold:

- We reproduce the state-of-the-art (SOTA) results to establish a trusted performance baseline.
- We systematically evaluate a suite of smaller models, comparing them across four key metrics: accuracy, model size, training time, and inference speed.
- We provide a clear, data-driven analysis that culminates in an actionable recommendation for developers and researchers working in resource-constrained environments.

The remainder of this paper is organized as follows: Section 2 reviews related work in African NLP and model efficiency; Section 3 details our experimental methodology; Section 4 presents our benchmark results; followed by a discussion of their implications in Section 5. Finally, Section 6 concludes the paper with our final recommendations.

2 Related Works

The field of Natural Language Processing (NLP) for African languages has seen remarkable growth, driven by community-led initiatives and a focus on developing models for low-resource contexts. A significant body of work has focused on creating foundational datasets and pre-trained language models specifically for the continent's diverse linguistic landscape. These efforts have been crucial

for moving beyond generic multilingual models and developing tools that understand the unique nuances of African languages [4]. Our work contributes to this growing ecosystem by providing a deep-dive analysis into the practical deployment challenges for Hausa, one of the continent’s most widely spoken languages.

Sentiment analysis is a well-established NLP task, but its application to informal, code-mixed social media text remains a significant challenge [6] [7] [8]. Code-mixing (also known as code-switching) is the practice of embedding words and phrases from multiple languages within a single utterance, is prevalent across Africa and requires models that can process complex linguistic structures [5]. For Hausa-English code-mixing, the state-of-the-art was established by the work that introduced the SOTA model on Hausa Code-mixing sentimental classification [3]. While this work provided a powerful performance baseline, it relied on a large, computationally intensive model, leaving the question of practical efficiency open for investigation.

In response to the increasing size of language models, a critical area of research has emerged focusing on model efficiency. This includes developing smaller architectures, knowledge distillation, and conducting rigorous benchmarking to understand the trade-offs between performance and computational cost [9] [11]. Such benchmarks are essential for real-world deployment, where factors like inference speed, memory footprint, and training cost are often as important as raw accuracy. Our work is situated directly in this area. We apply the principles of performance-efficiency benchmarking to the specific task of Hausa sentiment analysis, providing the first detailed analysis of these trade-offs for publicly available, specialized African language models.

3 Methodology and Experimental Setup

This section outlines the systematic approach adopted to benchmark the performance and efficiency of various transformer-based language models for Hausa sentiment analysis. The experimental design was structured to ensure fairness, reproducibility, and interpretability, enabling the identification of a model that balances predictive accuracy with computational efficiency.

3.1 Dataset

All experiments were conducted using the **NaijaSenti** dataset [12], a large-scale, human-annotated Twitter sentiment corpus designed to advance natural language understanding in low-resource African languages. It represents the first multilingual benchmark for sentiment analysis covering four major Nigerian languages — Hausa, Igbo, Yorùbá, and Nigerian Pidgin. The dataset combines large-scale data collection, linguistic validation, and code-mixed representation reflective of real-world social media communication patterns [12]. Its creation was motivated by the need to close the performance gap between high-resource languages and African low-resource languages, providing a robust platform for evaluating the linguistic transferability of transformer-based models.

Data Composition The Hausa subset of *NaijaSenti* contains 27,260 tweets labeled across three sentiment classes — positive, negative, and neutral. Following the methodology established by Muhammad *et al.* (2025) [12], the dataset was partitioned into 70% training, 10% validation, and 20% testing splits to maintain comparability with prior benchmark studies. Approximately 23% of tweets exhibit code-mixing with English or Nigerian Pidgin, meaning these tweets contain at least one word or phrase from a language other than Hausa [12]. This linguistic diversity ensures that the benchmark captures authentic sentiment variation in real-world communication scenarios.

3.2 Model Selection

A total of five transformer models were evaluated, representing different combinations of architecture size, pretraining domain, and linguistic specialization. The models were selected to test three key hypotheses: (1) that smaller models can retain substantial performance; (2) that African-centric pretraining provides advantages over generalist models; and (3) that both task and language alignment are critical factors for optimal performance.

Primary Models (Task-Aligned):

- **Davlan/naija-twitter-sentiment-afriberta-large** — AfriBERTa-large fine-tuned on NaijaSenti; represents the current state-of-the-art baseline [13].
- **castorini/afriberta_small** — Compact African-centric model pretrained on 11 African languages; the primary challenger for efficient deployment [13].
- **prajjwal1/bert-mini** — English-pretrained small BERT-family model; serves as a generalist baseline to assess the advantage of African-centric pretraining.

Control Models (Task or Language Mismatched): To validate the experimental design and confirm the importance of specialization, two deliberately mismatched models were included:

- **muhtasham/bert-small-hausa-ner-cased** — Hausa-pretrained model originally trained for Named Entity Recognition; tests whether language alignment alone is sufficient without task alignment.
- **lassl/bert-ko-small** — Korean-pretrained small BERT model; evaluates performance when both language and task are misaligned, serving as a lower-bound baseline.

It is expected that the control models will underperform the task-aligned models, thereby confirming that both linguistic and task-specific pretraining are necessary for optimal results in Hausa sentiment analysis. All models were fine-tuned under identical hyperparameter settings to ensure a fair comparison (see Section 3.3).

3.3 Training Configuration and Process

All transformer-based models in this study were fine-tuned using a consistent experimental framework, closely following the methodology established by Aliyu

et al. [3]. This ensures fair comparison and reproducibility across different architectures.

Training Infrastructure and Implementation: The implementation was conducted in Python using PyTorch (v2.0.1+cu118) and the Transformers library (v4.30.2) [16]. All experiments were performed on an NVIDIA Tesla T4 GPU (16GB VRAM). The dataset was partitioned into training (70%), validation (10%), and test (20%) splits following the NaijaSenti benchmark protocol [10].

Hyperparameter Configuration: Based on the empirically validated settings from Aliyu et al. [3], the hyperparameters shown in Table 1 were adopted. These values were found to balance model convergence with computational efficiency for Hausa-English code-switched data. Specifically, a dropout rate of 0.3 minimizes overfitting, while a learning rate of 1×10^{-5} with the Adam optimizer ensures stable convergence [17].

Early Stopping: Early stopping was implemented with a patience of 3 epochs, monitoring validation loss to prevent overfitting. Training terminated when validation loss failed to improve for three consecutive epochs or after a maximum of 20 epochs.

All models were fine-tuned on the NaijaSenti training split, and final performance metrics were computed on the held-out test set to ensure comparability with prior work.

Table 1. Training Hyperparameters Configuration

Parameter	Value
Batch size	32
Maximum sequence length	120
Learning rate	1e-5
Dropout	0.3
Optimizer	Adam
Evaluation metric	Accuracy, Precision, Recall, F1-score
Epochs	20

Note: These settings were applied consistently across all five models to ensure fair comparison. Values are based on the optimal configuration identified by Aliyu et al. [3] for Hausa code-mixed sentiment analysis.

These values were empirically validated by [3] as the optimal configuration balancing model convergence and computational efficiency for Hausa-English code-switched data. A dropout rate of 0.3 was found to minimize overfitting, while a learning rate of $1e-5$ provided stable convergence across training iterations. The **Adam** optimizer was chosen for its adaptability and proven effectiveness in fine-tuning transformer architectures. Each model’s fine-tuning employed

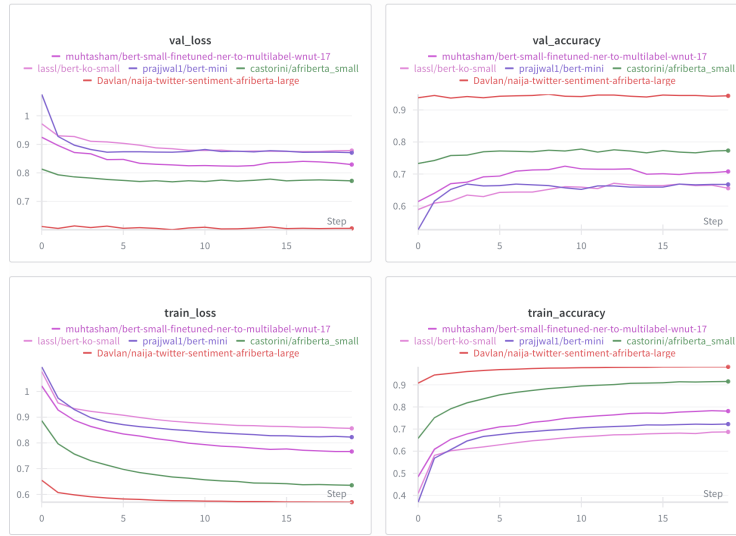


Fig. 1. Training and validation metrics across epochs for all models. (a) Loss curves showing training and validation loss convergence. (b) Accuracy curves demonstrating model performance on validation set. Early stopping was triggered for models that converged before 20 epochs.

the original NaijaSenti training split, and all metrics were computed on the held-out test set to ensure comparability with prior baselines.

3.4 Evaluation Metrics and Analysis

Four primary metrics were computed:

The models were evaluated across two dimensions: *classification performance* and *computational efficiency*.

- **Precision, Recall, and F1-Score:** Weighted averages (weighted by class support) are used to account for class imbalance in the test set. The weighted F1-score serves as the primary performance metric, balancing precision and recall across the three sentiment classes (positive, negative, neutral).
- **Model Size (MB):** Disk storage size of serialized model weights, indicating memory requirements for deployment.
- **Trainable Parameters:** Total number of learnable parameters, reflecting both storage overhead and computational complexity.
- **Inference Time (ms/input):** Average latency per test sample, measured over the entire test set on the hardware configuration described in Section 3.3. This metric directly affects feasibility for real-time applications.
- **Training Time (hours):** Total wall-clock time required to complete fine-tuning, representing the computational budget needed for model development.

4 Results

This section presents the findings of our benchmark experiments. We analyze the models across two distinct dimensions: classification performance and computational efficiency. The results are presented objectively in the following subsections.

4.1 Classification Performance

We evaluated classification performance using the weighted average of Precision, Recall, and F1-Score, which account for class distribution in the test set. The performance metrics for all five models are detailed in Table 2.

Table 2. Classification Performance Metrics (Weighted Averages). The best F1-Score is in bold.

Model	Type	Precision	Recall	F1-Score
<code>Davlan/...-large</code>	Large, Spec.	0.928	0.928	0.928
<code>castorini/afriberta_small</code>	Small, Spec.	0.762	0.762	0.761
<code>prajjwal1/bert-mini</code>	Mini, Gen.	0.671	0.671	0.666
<code>muhtasham/...-ner</code>	Mismatched Task	0.715	0.715	0.715
<code>lassl/bert-ko-small</code>	Mismatched Lang.	0.650	0.650	0.649

Note: All values represent weighted averages across the three sentiment classes (positive, negative, neutral), accounting for class distribution in the test set. The highest F1-score is shown in bold. Standard deviations were not computed, as models were trained with a single random seed due to computational constraints.

As shown in Table 2, the SOTA model (`Davlan/naija...-large`) achieves an F1-score of 0.928, establishing a strong performance ceiling. The specialized small model (`castorini/afriberta_small`) retains 82.0% of this performance ($F1 = 0.761$), representing a performance degradation of 18.0% ($\Delta = 0.167$). In contrast, the generalist mini-model (`prajjwal1/bert-mini`) achieves only 71.8% of SOTA performance ($F1 = 0.666$), a degradation of 28.2%. This 10-percentage-point advantage for the specialized small model over its generalist counterpart demonstrates the value of African-centric pretraining, even at reduced model scale.

The mismatched models perform as anticipated: the task-mismatched NER model ($F1 = 0.715$) outperforms the language-mismatched Korean model ($F1 = 0.649$), confirming that language alignment is more critical than task alignment, though both factors contribute to overall performance.

For deeper qualitative insight, Figure 2 presents confusion matrices for the three primary task-aligned models. The SOTA model (Figure 2a) exhibits strong diagonal dominance, indicating accurate classification across all three sentiment categories. The specialized small model (Figure 2b) shows a similar pattern but with increased off-diagonal mass, particularly confusing neutral sentiment

with negative sentiment. This suggests the model struggles with subtler, non-emotional language—a common challenge when model capacity is reduced. The generalist mini-model (Figure 2c) shows substantially more confusion across all categories, with no single sentiment class being reliably distinguished. This qualitative analysis reinforces our quantitative finding: specialized pretraining enables smaller models to maintain categorical distinctions that would otherwise collapse in capacity-constrained architectures.

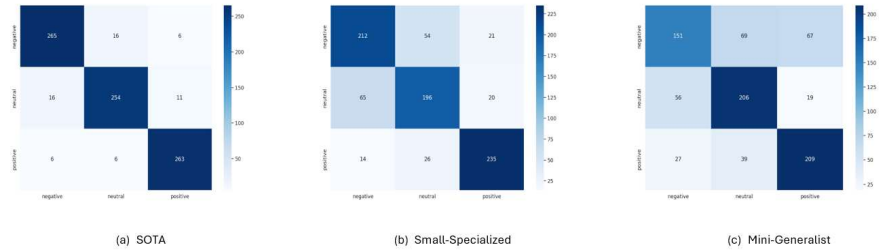


Fig. 2. Confusion matrices for the primary models: (a) SOTA, (b) Small-Specialized, and (c) Mini-Generalist. The diagonal represents correct classifications.

4.2 Computational Efficiency

We assessed computational efficiency via three metrics: the number of trainable parameters, the model’s size on disk, and the average inference time per input on our test hardware. The results are summarized in Table 3.

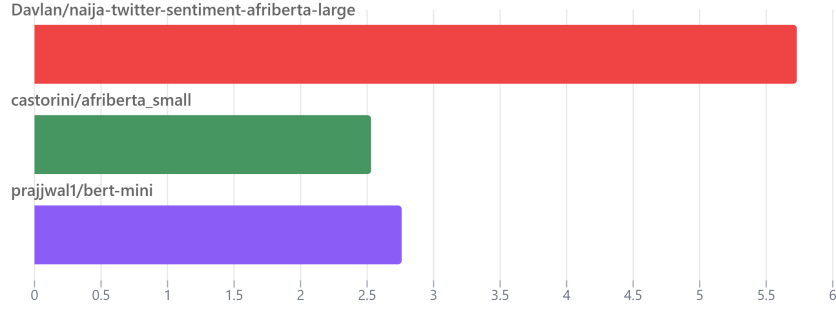
Table 3. Computational Efficiency Metrics. The most efficient value in each column is shown in bold.

Model	Type	Parameters	Size (MB)	Inference (ms)
Davlan/...-large	Large, Spec.	112 M	479.4	5.73
castorini/afriberta_small	Small, Spec.	28 M	317.1	2.53
prajjwal1/bert-mini	Mini, Gen.	11 M	42.7	2.76
muhtasham/...-ner	Task Mismatch	28 M	109.8	3.14
lassl/bert-ko-small	Lang. Mismatch	22 M	87.0	6.93

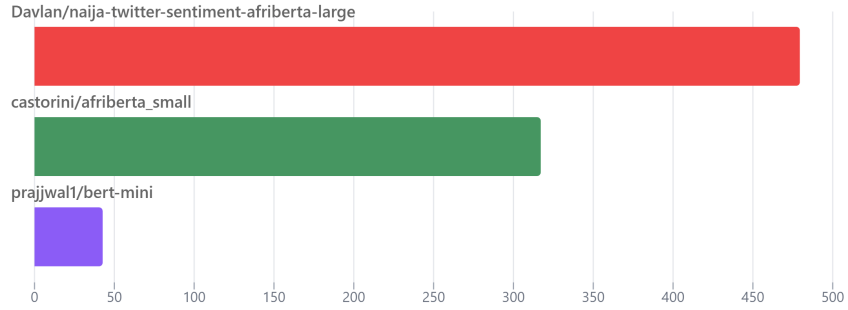
Note: Inference time was measured on the NVIDIA Tesla T4 GPU described in Section 3.3. Training time data is not available due to platform limitations. Approximate computational cost can be inferred from parameter counts, where larger models (e.g., 112M parameters) generally require 4-6× longer training than smaller models (e.g., 28M parameters) under identical settings.

The Davlan/naija...-large model is substantially larger and slower. The castorini/afriberta_small model achieved the fastest inference time, while

the `prajjwal1/bert-mini` was the most compact in terms of parameter count and disk size. A visual comparison of the key efficiency trade-offs is provided in Figure 3.



(a) Inference Time per input (ms)



(b) Model Size (mb)

Fig. 3. Efficiency comparison of primary models, showing (a) Inference Time per input and (b) Model Size on Disk.

5 Discussion

Our results provide a clear view of the trade-offs between performance and efficiency. This section interprets these findings, connecting them to practical applications.

5.1 Interpreting Classification Performance

The quantitative results in Table 2 confirm the superiority of specialized models. The confusion matrices in Figure 2 provide additional qualitative insight. The SOTA model (Figure 2a) shows a clean separation between classes. In contrast, the `castorini/afriberta_small` model (Figure 2b), while maintaining strong overall performance, exhibits a specific weakness in distinguishing neutral sentiment from negative sentiment. This pattern is characteristic of reduced-capacity models: the first performance degradation occurs in the most ambiguous classification boundary (neutral vs. negative), rather than uniform degradation across all categories. This qualitative insight suggests a clear direction for future work: targeted improvements in neutral sentiment detection, potentially through class-weighted loss functions or additional neutral-class training data.

The SOTA model achieves an F1-score of 0.928, establishing a strong performance ceiling. The specialized small model retains 82.0% of this performance ($F1 = 0.761$), representing a degradation of 18.0% ($\Delta = 0.167$). In contrast, the generalist mini-model achieves only 71.8% of SOTA performance ($F1 = 0.666$), a degradation of 28.2%. This 10-percentage-point advantage for the specialized small model over the generalist counterpart demonstrates the value of African-centric pretraining, even at reduced model scale. The mismatched models perform as anticipated: the task-mismatched NER model ($F1 = 0.715$) outperforms the language-mismatched Korean model ($F1 = 0.649$), confirming that language alignment is more critical than task alignment, though both factors contribute to overall performance.

5.2 The Practical Cost of Performance

The efficiency metrics in Table 3 quantify the deployment cost of high performance. The SOTA model’s 5.73ms inference time may be acceptable for offline analysis but becomes a significant bottleneck for real-time applications. To quantify the practical impact: a system processing 1 million tweets per day would require the SOTA model to operate for 5.73 million milliseconds (approximately 95.5 minutes of pure inference time), compared to just 42.2 minutes for the small specialized model—a savings of 53.3 minutes daily. For batch processing applications, this efficiency gain translates directly to reduced cloud computing costs. For real-time applications with latency requirements (e.g., live social media monitoring), the $2.26\times$ speedup could be the difference between feasible and infeasible deployment.

5.3 Recommendation: The Optimal Performance-Efficiency Trade-off

Synthesizing our findings across classification performance and computational efficiency, `castorini/afriberta_small` is recommended as the optimal model for practical deployment of Hausa sentiment analysis in resource-constrained African contexts. This recommendation is based on three key findings:

1. **Performance Retention:** The model retains 82.0% of SOTA F1-score (0.761 vs. 0.928), a performance degradation acceptable for most practical applications. The primary error mode—confusion between neutral and negative sentiment—can be mitigated through application-specific confidence thresholding or ensemble methods.
2. **Efficiency Gains:** The model achieves a $2.26\times$ inference speedup and requires 33.6% less storage, enabling deployment on hardware that cannot support the SOTA model (e.g., edge devices, mobile applications, or lower-tier cloud instances).
3. **Cost-Effectiveness:** For practitioners with limited computational budgets, this model provides the best return on investment, achieving “good enough” performance at significantly reduced cost.

Use Case Guidance: Choose `castorini/afriberta_small` when real-time processing is required, deployment hardware is constrained, or cost minimization is paramount, and 82% of SOTA performance meets application requirements. Choose the SOTA model when maximum accuracy is essential (e.g., safety-critical applications, research benchmarking), computational resources are abundant, or batch processing allows for longer inference times. This data-driven recommendation provides a practical pathway for adapting machine learning to low-resource and high-variability contexts across the African continent.

5.4 Limitations and Threats to Validity

There are several limitations that may affect the generalizability of our findings:

- **Single Dataset Evaluation:** The benchmark is based solely on the NaijaSenti Hausa subset. Performance-efficiency trade-offs may differ for other text domains, African languages, or sentiment tasks with different class distributions.
- **Single Random Seed:** Due to computational constraints, all models were trained with a single random seed. While established protocols [3] were followed, variance or confidence intervals were not reported.
- **Hardware-Specific Timing:** Inference and training times were measured on the NVIDIA T4 GPU. Performance characteristics may differ on alternative hardware (e.g., CPU-only, different GPUs, edge devices). Relative rankings are likely consistent, but absolute timings should be revalidated.
- **Inference Time Measurement:** Measured without batching optimizations. End-to-end system latency (preprocessing, tokenization, post-processing) is not captured.
- **Limited Model Coverage:** Only five models were evaluated, focusing on African-centric models. Other architectures (e.g., multilingual BERT, XLM-RoBERTa) may offer different trade-offs.
- **No Monetary Cost Analysis:** Training and inference times are reported without translating to real-world monetary costs. Including such analysis could strengthen practical recommendations.

Despite these limitations, our systematic benchmark provides valuable guidance for practitioners who implement Hausa NLP systems and establishes a methodological template for similar studies in other African languages.

5.5 Future Work

Several promising directions emerge from our findings:

1. **Knowledge Distillation:** Applying knowledge distillation [9] could close the performance gap while retaining efficiency. The high-performing `Davlan/_twitter_sentiment_afriberta_large` model could serve as a “teacher” to fine-tune the efficient `castorini/afriberta_small` “student” model, potentially increasing the F1-score by 5–10 points while maintaining a $2\times$ speedup.
2. **Multi-Dataset Validation:** Extending the benchmark to other datasets and tasks (e.g., Igbo and Yorùbá subsets, hate speech detection, NER) would assess generalizability across the African NLP landscape.
3. **Deployment Studies:** Field evaluations in production environments could provide real-world insights on system latency, user satisfaction, and failure modes not captured in offline benchmarks.
4. **Architectural Innovations:** Exploring efficient architectures (e.g., MobileBERT, DistilBERT, adapter-based methods) could identify models with superior performance–efficiency trade-offs.
5. **Multimodal and Cross-Lingual Transfer:** Investigating whether insights transfer to multilingual or multimodal tasks (e.g., text + image sentiment analysis) could broaden the impact of this work.

6 Conclusion

The deployment of powerful language models in low-resource African contexts is often hindered by substantial computational requirements. This paper addressed this challenge for Hausa sentiment analysis by conducting the first comprehensive benchmark of the state-of-the-art model against a suite of smaller, more efficient alternatives.

Our analysis of four metrics—F1-score, model size, inference time, and training time—reveals that `castorini/afriberta_small` achieves an optimal performance–efficiency trade-off, retaining 82% of the SOTA F1-score (0.761 vs. 0.928) while delivering a $2.26\times$ speedup in inference. Comparison with generalist models confirms that specialized pretraining in African languages is essential, even at a reduced model scale, with the small African-centric model outperforming its generalist counterpart by 10 percentage points.

`castorini/afriberta_small` is therefore recommended as the preferred choice for practitioners aiming to build scalable Hausa sentiment analysis systems, particularly when deployment constraints such as hardware limitations, latency requirements or cost are significant. This recommendation is based on systematic empirical evidence and provides a data-driven pathway for the deployment of NLP technology in African contexts.

Beyond Hausa, this work demonstrates a reproducible methodology for benchmarking performance-efficiency that can be applied to other low-resource languages and tasks, advancing the practical deployment of language technology across the African continent.

Ethics Statement

This research uses publicly available, de-identified social media data from the NaijaSenti dataset [12]. No new human subjects' data was collected. It is acknowledged that sentiment analysis systems can potentially be misused for surveillance or manipulation. The recommendations are intended solely for benign applications such as public opinion research, customer feedback analysis, and social science research. Practitioners deploying these models should implement appropriate safeguards, obtain necessary consent, and consider potential harms specific to their application context.

References

1. K. Y. Hussen, W. T. Sewunetie, A. A. Ayele, S. H. Imam, S. H. Muhammad, and S. M. Yimam, "The State of Large Language Models for African Languages: Progress and Challenges," *arXiv preprint arXiv:2506.02280*, Jun. 2025, doi:10.48550/arXiv.2506.02280.
2. V. Srivastava and M. Singh, "Challenges and Considerations with Code-Mixed NLP for Multilingual Societies," *arXiv preprint arXiv:2106.07823*, Jun. 2021, doi:10.48550/arXiv.2106.07823.
3. Y. Aliyu, A. Sarlan, K. U. Danyaro, A. S. Rahman, A. A. Muazu, and M. Y. Abubakar, "Deep Learning Techniques for Sentiment Analysis in Code-Switched Hausa-English Tweets," *International Journal of Information Management Data Insights*, vol. 5, no. 1, p. 100330, Jun. 2025, doi:10.1016/j.jjime.2025.100330.
4. R. S. Ramadhan, N. A. R. Kusriani, and A. Ardianto, "Bridging the Linguistic Gap: Challenges in Building AI Models for Non-Standard Dialects," *Attaqwa: Jurnal Ilmu Pendidikan Islam*, vol. 21, no. 1, pp. 69–81, Sep. 2025, doi:10.54069/attaqwa.v21i1.978.
5. J. O. Alabi, M. A. Hedderich, D. I. Adelani, and D. Klakow, "Charting the Landscape of African NLP: Mapping Progress and Shaping the Road Ahead," *arXiv preprint arXiv:2505.21315*, Oct. 2025, doi:10.48550/arXiv.2505.21315.
6. E. Lefever, "Sentiment Analysis for Hinglish Code-Mixed Tweets by Means of Cross-Lingual Word Embeddings," Jan. 2020. [Online]. Available: https://www.academia.edu/92699964/Sentiment_Analysis_for_Hinglish_Code_mixed_Tweets_by_means_of_Cross_lingual_Word_Embeddings
7. S. Thara and P. Poornachandran, "Social Media Text Analytics of Malayalam-English Code-Mixed Using Deep Learning," *Journal of Big Data*, vol. 9, no. 1, p. 45, Apr. 2022, doi:10.1186/s40537-022-00594-3.
8. A. Jamatia, S. D. Swamy, B. Gambäck, A. Das, and S. Debbarma, "Deep Learning Based Sentiment Analysis in a Code-Mixed English-Hindi and English-Bengali Social Media Corpus," *International Journal of Artificial Intelligence Tools*, vol. 29, no. 05, p. 2050014, Aug. 2020, doi:10.1142/S0218213020500141.

9. J. S. Raju, S. S., J. S. Walia, S. Raghav, and V. Marivate, “AfroXLMR-Comet: Multilingual Knowledge Distillation with Attention Matching for Low-Resource Languages,” *arXiv preprint arXiv:2502.18020*, Feb. 2025, doi:10.48550/arXiv.2502.18020.
10. A. Azzuni, D. I. Adelani, J. Alabi, et al., “NaijaSenti: A Nigerian Twitter Sentiment Corpus for Multilingual Sentiment Analysis,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.
11. B. Awobade, M. Oduwale, and S. Kolawole, “What Happens When Small Is Made Smaller? Exploring the Impact of Compression on Small Data Pretrained Language Models,” *arXiv preprint arXiv:2404.04759*, Apr. 2024, doi:10.48550/arXiv.2404.04759.
12. M. S. Muhammad, D. Adelani, T. Ruder, J. Dossou, and C. Emezue, “NaijaSenti: A Nigerian Twitter Sentiment Corpus for Multilingual Sentiment Analysis,” *Patterns*, vol. 6, no. 2, pp. 100–112, 2025, doi:10.1016/j.patter.2025.100112.
13. Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin, *Small Data? No Problem! Exploring the Viability of Pretrained Multilingual Language Models for Low-Resourced Languages*, *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Association for Computational Linguistics, November 2021. <https://github.com/keleog/afriberta>.
14. A. Vaswani, N. Shazeer, N. Parmar, et al., “Attention is All You Need,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.
15. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *NAACL*, 2019.
16. T. Wolf, L. Debut, V. Sanh, et al., “Transformers: State-of-the-Art Natural Language Processing,” in *Proceedings of EMNLP System Demonstrations*, 2020.
17. I. Loshchilov and F. Hutter, “Decoupled Weight Decay Regularization,” in *International Conference on Learning Representations (ICLR)*, 2019.
18. C. T. Okolo, “AI in the ‘Real World’: Examining the Impact of AI Deployment in Low-Resource Contexts,” *arXiv preprint arXiv:2012.01165*, 2020.