Cross-lingual transfer of multilingual models on low resource African Languages

Harish Thangaraj[1], Ananya Chenat[1], Jaskaran Singh Walia[1], and Vukosi Marivate[2,3]

[1] Vellore Institute of Technology, India
[2] Data Science for Social Impact, University of Pretoria, South Africa
[3] Lelapa AI

Abstract. Cross-lingual transfer learning is crucial for developing NLP technologies for low-resource African languages, yet the optimal modeling strategy remains an open question. This paper presents a benchmark for cross-lingual transfer from Kinyarwanda to Kirundi, two closely related Bantu languages. We evaluate the performance of traditional monolingual architectures (BiGRU, CNN, Char-CNN) against three distinct multilingual transformer models: the global mBERT, the Africa-centric AfriBERT, and the language-family-specific BantuBERTa. Our evaluation covers both zero-shot and fine-tuned news classification, and critically, we measure the degree of catastrophic forgetting on the source language after fine-tuning. Our results demonstrate that the regionally-focused AfriBERT achieves the highest cross-lingual accuracy (88.3%) after fine-tuning. Furthermore, we find that large-scale pre-training is essential for robustness; mBERT and AfriBERT exhibit minimal forgetting, while BantuBERTa and the traditional models suffer a severe performance degradation. This study highlights the importance of regionally focused multilingual models for transfer between related African languages and establishes catastrophic forgetting as a critical evaluation metric for such tasks.

Keywords: Cross-lingual Transfer · Low-resource Language Modelling · Multilingual Models · Catastrophic Forgetting · Kinyarwanda and Kirundi · News Classification

1 Introduction

The development of robust Natural Language Processing (NLP) technologies for the world's diverse languages remains a significant challenge, particularly for the over 2000 languages spoken across Africa. Due to data scarcity, cross-lingual transfer learning—where learnings from a relatively higher-resource language is adapted for a lower-resource one—has become an essential strategy. The effectiveness of this transfer, however, depends critically on the choice of model architecture and pre-training regimen.

A central debate revolves around the optimal approach for transfer. On one hand, traditional monolingual models (e.g., CNN, BiGRU) can be trained to capture the specific nuances of a source language. On the other, large multilingual transformer models like mBERT [4] offer powerful, pre-trained representations learned from a global

corpus. More recently, regionally-focused models such as AfriBERT [14] and BantuBERTa [16,17] have emerged, promising more specialised transfer capabilities by pre-training on curated sets of related languages. Yet, there is a lack of comprehensive benchmarks that compare these different strategies head-to-head in a realistic African language context, particularly regarding the crucial trade-off between transfer performance and the risk of catastrophic forgetting—where a model loses its proficiency in the original source language after fine-tuning.

This paper addresses this gap by conducting a rigorous benchmarking study of cross-lingual transfer from Kinyarwanda to Kirundi, two closely related Bantu languages. Extending the work of [13], we compare traditional neural architectures against global, regional, and language-family-specific multilingual transformer models. Our core research question is: Which transfer learning strategy yields the best performance on the target language while retaining performance on the source language?

Our main contributions are:

1. A comprehensive benchmark of six models (mBERT, AfriBERT, BantuBERTa, BiGRU, CNN, Char-CNN) for Kinyarwanda-to-Kirundi news classification, evaluated in both zero-shot and fine-tuned settings.
2. An empirical analysis of catastrophic forgetting, quantifying how different model architectures are affected by fine-tuning on a low-resource target language.
3. A clear demonstration of the superiority of regionally-focused multilingual models (AfriBERT) for this task, providing practical guidance for researchers and practitioners working on related African languages.

2 Related Work

Our research is situated at the intersection of cross-lingual transfer learning, model architecture, and Natural Language Processing for low-resource African languages.

2.1 Cross-Lingual Transfer for Low-Resource Languages

Addressing data scarcity in low-resource languages (LRLs) is a central challenge in modern NLP [10]. Cross-lingual transfer learning has emerged as a dominant strategy, where learnings from a higher-resource source language is transferred to a lower-resource target language. Early approaches demonstrated the viability of transfer using techniques like annotation projection and direct transfer with LSTM architectures, showing that even single-source transfer from a distant language like English can improve performance [20]. More recent work has focused on improving transfer efficiency through data augmentation [19], back-translation [7], and sophisticated adapter-based architectures that leverage multi-source information [18]. Our work builds on the "direct transfer" approach, focusing on a scenario where the source and target languages are closely related.

2.2 Paradigms in Transfer: Monolingual vs. Multilingual Models

A key debate in cross-lingual transfer concerns the optimal model architecture. One paradigm focuses on transferring learnings from powerful **monolingual models**. Studies have shown that representations from a single-language BERT can be effectively transferred to other languages, preserving both syntactic and semantic knowledge [6]. Monolingual models fine-tuned on specific dialects, such as DarijaBERT for Moroccan Arabic, have even been shown to outperform larger multilingual models on specific tasks, highlighting their ability to capture fine-grained linguistic nuances [2].

The alternative paradigm leverages large **multilingual models** like mBERT [4] and XLM-R [3], which learn a shared representation space across many languages during pre-training. These models have demonstrated strong zero-shot and few-shot cross-lingual capabilities [21]. Their effectiveness often depends on aligning latent representations between languages, for which methods like unsupervised machine translation have been explored [5]. However, the performance of these global models can be surpassed by regionally-focused multilingual models, such as ARBERT and MARBERT for Arabic, which achieve state-of-the-art results by pre-training on a curated set of related languages and dialects [1]. Our paper directly contributes to this debate by empirically comparing the cross-lingual performance of both monolingual and various multilingual models in a resource-constrained setting.

2.3 Advancements in NLP for African Languages

There is a growing body of work dedicated to developing NLP resources and models specifically for African languages. Research has shown that multilingual training incorporating related languages significantly enhances performance for languages like isiZulu and Sepedi [11], and multilingual neural machine translation models consistently outperform single-pair models for African languages [9].

This has led to the development of Africa-centric language models. Studies have demonstrated that multilingual models pre-trained on African languages outperform monolingual baselines in transfer tasks [15] [12]. This motivates our use of **AfriBERT** [14], which is pre-trained on 11 African languages, and **BantuBERTa** [16,17], which focuses specifically on the Bantu language family. Our work serves as a direct extension of [13], which introduced the KINNEWS and KIRNEWS datasets and provided initial benchmarks using traditional architectures like BiGRU and CNN. By including and benchmarking modern, Africa-centric transformer models against these baselines, we provide an updated and more comprehensive analysis of cross-lingual transfer between Kinyarwanda and Kirundi.

3 Experiments

We conduct a series of experiments to benchmark cross-lingual transfer from Kinyarwanda to Kirundi. Our methodology is organised into four parts: the datasets used, the models evaluated, the experimental training protocol, and the evaluation metrics.

3.1 Datasets and Preprocessing

The study utilises the KINNEWS (Kinyarwanda) and KIRNEWS (Kirundi) datasets introduced by [13]. These two languages belong to the Rwanda-Rundi dialect continuum and share significant lexical and grammatical similarities, making them ideal candidates for a cross-lingual transfer study.

The raw data was preprocessed by removing duplicate articles and irrelevant metadata. For a unified text representation, the 'title' and 'content' of each news article were concatenated. The datasets were filtered to include only the 12 semantic categories common to both languages. A summary of the resulting datasets is provided in Table 1.

Table 1: Statistics of the preprocessed KINNEWS and KIRNEWS datasets used in this study.

Language	Dataset	Articles	Classes	Train/Test Split
Kinyarwanda	KINNEWS	21,268	12	17,014 / 4,254
Kirundi	KIRNEWS	4,612	12	3,690 / 922

3.2 Models

We evaluate two classes of models to compare their cross-lingual transfer capabilities.

Multilingual Transformer Models. We use three pre-trained transformer-based models known for their multilingual capabilities:

- **mBERT** [4]: A multilingual BERT model pre-trained on Wikipedia text from 104 languages, serving as a strong general-purpose cross-lingual baseline.
- **AfriBERT** [14]: A model pre-trained on a curated corpus of 11 African languages, including Kinyarwanda and Kirundi, designed to better capture regional linguistic nuances.
- **BantuBERTa** [16,17]: A model pre-trained specifically on Bantu languages, intended to leverage the typological similarities within this language family.

Traditional Neural Models. As monolingual baselines, we adapt three standard neural architectures from [13], which rely on language-specific embeddings:

- **CNN**: A Convolutional Neural Network for text classification as proposed by [8].
- **Char-CNN**: A character-level CNN, capable of capturing sub-word morphological features.
- **BiGRU**: A Bidirectional Gated Recurrent Unit network, which processes text sequentially to capture contextual information.

For these models, we trained a Word2Vec model (skip-gram, vector size 50, window size 5) on the Kinyarwanda training corpus to generate input word embeddings.

3.3 Experimental Design and Training

Our experimental protocol involves a three-stage process to comprehensively evaluate cross-lingual transfer and catastrophic forgetting.

1. **Source Language Training:** All models are first trained on the Kinyarwanda training set for the 12-class news classification task which was split into 90% for training and 10% for validation..
2. **Target Language Evaluation:** The Kinyarwanda-trained models are then evaluated on the Kirundi test set in two distinct settings:
   - *Zero-Shot Transfer:* The model is evaluated directly on the Kirundi test set without any further training.
   - *Fine-Tuned Transfer:* The model is fine-tuned on the Kirundi training set before being evaluated on the Kirundi test set.
3. **Forgetting Evaluation:** After fine-tuning on Kirundi, the models are re-evaluated on the Kinyarwanda test set to measure the degradation in source-language performance.

The transformer models were trained using the hyperparameters detailed in Table 2 and the Traditional models were adopted from [13]

Table 2: Hyperparameters for fine-tuning the transformer models.

Parameter	**Value**
Number of Labels	12
Input Sequence Length	128
Truncation	True
Padding	True
Device	`MPS`
Number of Training Epochs:	
mBERT	8
AfriBERT	25
BantuBERTa	8
Per-Device Train/Eval Batch Size	32
Warmup Steps	500
Weight Decay	0.01
Logging Steps	10
Load Best Model at End	True
Metric for Best Model	`f1`
Evaluation Strategy	`steps`

All experiments were repeated five times with different random seeds, and we report the mean of accuracy and F1-score with least standard deviations.

3.4 Evaluation Metrics

We use three metrics to assess model performance across different stages of the experiment:

- **Accuracy:** The proportion of correctly classified examples.
- **F1-Score:** The macro-averaged F1-score, which is the harmonic mean of precision and recall, providing a balanced measure for multi-class classification.
- **Forgetting:** The percentage decrease in performance on the original (Kinyarwanda) task after the model has been fine-tuned on the new (Kirundi) task. It is calculated as:

$$\text{Forgetting } (\%) = \left(\frac{\text{Acc}_{\text{initial}} - \text{Acc}_{\text{final}}}{\text{Acc}_{\text{initial}}} \right) \times 100 \tag{1}$$

  where $\text{Acc}_{\text{initial}}$ is the accuracy on Kinyarwanda before fine-tuning on Kirundi, and $\text{Acc}_{\text{final}}$ is the accuracy after.

4 Results

This section presents the empirical results of our cross-lingual transfer experiments. We report on three key aspects: (1) the baseline performance of models on the source language (Kinyarwanda), (2) the cross-lingual performance on the target language (Kirundi) before and after fine-tuning, and (3) the extent of catastrophic forgetting on the source language after fine-tuning.

4.1 Cross-Lingual Transfer Performance

The core results for the Kinyarwanda-to-Kirundi transfer task are presented in Table 3. In the zero-shot setting (Before FT), the multilingual models demonstrated a clear advantage, with BantuBERTa achieving the highest accuracy (74.5%). In contrast, the traditional models performed near random chance. After fine-tuning on Kirundi data (After FT), all models showed substantial gains. AfriBERT emerged as the top-performing model in terms of accuracy, reaching **88.3%**. Notably, the BiGRU model achieved the highest F1-score at **87.9%**, highlighting its strength as a traditional baseline.

Table 3: Metrics describing cross-lingual testing on Kirundi

Model	Accuracy before FT	F1 before FT	Accuracy after FT	F1 after FT
mBERT	0.5872	0.5917	0.8462	0.8422
AfriBERT	0.7421	0.7474	**0.8830**	**0.8787**
BantuBERTa	0.7454	0.7375	0.8657	0.8606
BiGRU	0.2404	0.2300	**0.8332**	**0.8790**
CNN	0.2190	0.2320	0.5913	0.5732
Char-CNN	0.1916	0.1621	0.4879	0.4764

4.2 Monolingual Performance and Catastrophic Forgetting

Table 4 details model performance on the source Kinyarwanda dataset. Before fine-tuning (Table 4a), the traditional models BiGRU and CNN achieved the highest monolingual accuracy at 88.5% and 87.4%, respectively. Table 4b shows the performance on Kinyarwanda after the models were fine-tuned on Kirundi. The large multilingual models, mBERT and AfriBERT, proved highly resilient, retaining most of their original performance with forgetting rates of just **3.0%** and 5.1%. In stark contrast, BantuBERTa and all traditional models suffered from severe catastrophic forgetting, with performance drops exceeding 70%. Figure 1 visually summarises this degradation.

Table 4: Performance on the Kinyarwanda test set. (a) Initial monolingual performance before any cross-lingual fine-tuning. (b) Performance after models were fine-tuned on Kirundi, with the calculated percentage of forgetting. Best results are in **bold**.

(a) Performance before fine-tuning.

Model	Accuracy	F1 score
mBERT	0.7884	0.7747
AfriBERT	0.8498	0.8447
BantuBERTa	0.8601	0.8555
BiGRU	**0.8851**	0.8434
CNN	0.8740	**0.8660**
Char-CNN	0.6930	0.6823

(b) Performance after fine-tuning on Kirundi.

Model	Accuracy	Forget %
mBERT	0.7645	**3.03**
AfriBERT	**0.8061**	5.14
BantuBERTa	0.2172	74.00
BiGRU	0.2329	73.68
CNN	0.2207	74.86
Char-CNN	0.1968	71.50

Figure 1 portrays a graphical representation of forgetting and improvement after fine-tuning, for Kinyarwanda.

5 Discussion

Our results offer several key insights into cross-lingual transfer for low-resource African languages. While baseline models achieve the highest overall F1 performance before fine-tuning, multilingual models demonstrate competitive results and exhibit clear advantages in retaining source-language performance after fine-tuning. This highlights the strength of pre-trained representations as a foundation that monolingual models trained from scratch on small datasets cannot match.

The standout performance of AfriBERT (88.3% accuracy) over mBERT (84.6%) is noteworthy. AfriBERT is trained on a focused set of 11 African languages, which allows it to learn representations that capture shared lexical, morphological, and syntactic patterns within this group. In contrast, mBERT is trained on 104 languages, and the uneven distribution of corpora across languages can reduce its ability to model low-resource languages effectively. These factors likely contribute to AfriBERT's superior

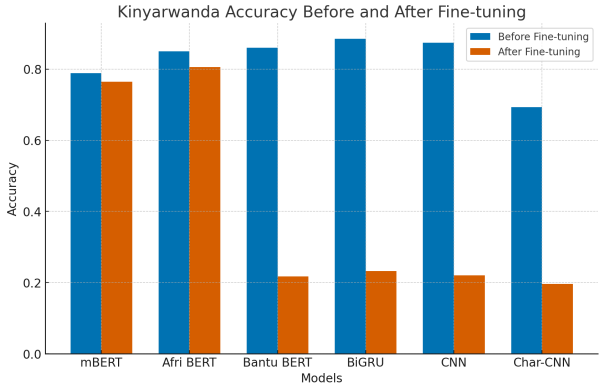

Fig. 1: Performance degradation on Kinyarwanda after fine-tuning, illustrating the forgetting gap between large multilingual models and others.

performance, indicating that limiting pretraining to a smaller, linguistically related set of languages can improve cross-lingual transfer, especially for Bantu languages.

critical finding is the large difference in catastrophic forgetting across models. AfriBERT and mBERT show low forgetting (5.1% and 3.0%), whereas BantuBERTa and the traditional models experience much higher forgetting (74.0% and near-total loss). We hypothesise that this is influenced by the scale and diversity of pre-training and also by tokenizer coverage: models with better tokenization for Kinyarwanda and Kirundi are likely to retain knowledge more effectively. In contrast, models with smaller pre-training or less suitable tokenizers are more prone to forgetting. This suggests that both pre-training scale and tokenization quality contribute to the observed differences.

Finally, the results underscore the value of linguistic similarity. The high degree of mutual intelligibility between Kinyarwanda and Kirundi provides an ideal scenario for transfer learning. Our work empirically demonstrates that modern transformer architectures are highly effective at leveraging this lexical and structural overlap, even with limited fine-tuning data.

6 Conclusion

This study benchmarked the cross-lingual transfer capabilities of multilingual and traditional neural models on a Kinyarwanda-to-Kirundi news classification task. We demonstrated that pre-trained multilingual models, particularly the regionally-focused AfriBERT, significantly outperform traditional architectures in transfer accuracy. Our analysis also quantified the critical challenge of catastrophic forgetting, revealing the superior resilience of models with larger pre-training corpora. The core contribution of this work is an empirical comparison that provides clear guidance on model selection for tasks involving closely related, low-resource languages, highlighting the strengths of regional multilingual models and the pitfalls of catastrophic forgetting.

7 Limitations

Our study, while providing valuable benchmarks, has several limitations that open avenues for future research. The scope of our analysis is confined to the Kinyarwanda-Kirundi language pair using news-domain datasets, and thus the findings may not directly generalise to other Bantu languages or text domains. A significant methodological limitation is the catastrophic forgetting observed, particularly in BantuBERTa and the traditional models, as our experimental setup did not employ mitigation strategies like continual learning; investigating such techniques is a critical next step. Finally, the performance variance between the multilingual models suggests that the scale and composition of the pre-training corpus are key. A deeper analysis of these pre-training regimens is needed to fully understand the drivers of robust cross-lingual transfer.

Acknowledgments. The authors gratefully acknowledge the support of the ABSA Chair of Data Science for facilitating this research. DSFSI is supported by the UK International Development and the International Development Research Centre, Ottawa, Canada as part of the AI for Development: Responsible AI, Empowering People Program (AI4D). DSFSI is thankful for gifts from NVIDIA, Google.org, OpenAI and Meta which enable our research.

References

1. Abdul-Mageed, M., Elmadany, A., Nagoudi, E.M.B.: ARBERT & MARBERT: Deep bidirectional transformers for Arabic. arXiv preprint arXiv:2101.01785 (2020)
2. Boudad, N., Faizi, R., Thami, R.O.H.: Cross-multilingual, cross-lingual and monolingual transfer learning for Arabic dialect sentiment classification (2023)
3. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8440–8451. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.747, https://aclanthology.org/2020.acl-main.747/
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference **1**, 4171–4186 (10 2018), https://arxiv.org/abs/1810.04805v2
5. Fei, H., Li, P.: Cross-lingual unsupervised sentiment classification with multi-view transfer learning. In: Proceedings of the 58th annual meeting of the association for computational linguistics (2020)
6. Gogoulou, E., Ekgren, A., Isbister, T., Sahlgren, M.: Cross-lingual transfer of monolingual models. 2022 Language Resources and Evaluation Conference, LREC 2022 pp. 948–955 (9 2021), https://arxiv.org/abs/2109.07348v2
7. Karakanta, A., Dehdari, J., van Genabith, J.: Neural machine translation for low-resource languages without parallel corpora. Machine Translation **32**, 167–189 (6 2018). https://doi.org/10.1007/S10590-017-9203-5/FIGURES/7, https://link.springer.com/article/10.1007/s10590-017-9203-5
8. Kim, Y.: Convolutional neural networks for sentence classification (2014), https://arxiv.org/abs/1408.5882

9. Lakew, S.M., Negri, M., Turchi, M.: Low resource neural machine translation: A benchmark for five African languages (3 2020), https://arxiv.org/abs/2003.14402v1
10. Magueresse, A., Carles, V., Heetderks, E.: Low-resource languages: A review of past work and future challenges (6 2020), https://arxiv.org/abs/2006.07264v1
11. Mesham, S., Hayward, L., Shapiro, J., Buys, J.: Low-resource language modelling of South African languages (4 2021), https://arxiv.org/abs/2104.00772v1
12. Muhammad, S.H., Abdulmumin, I., Ayele, A.A., Ousidhoum, N., Adelani, D.I., Yimam, S.M., Ahmad, I.S., Beloucif, M., Mohammad, S.M., Ruder, S., Hourrane, O., Brazdil, P., Jorge, A., Ali, F.D.M.A., David, D., Osei, S., Bello, B.S., Ibrahim, F., Gwadabe, T., Rutunda, S., Belay, T., Messelle, W.B., Balcha, H.B., Chala, S.A., Gebremichael, H.T., Opoku, B., Arthur, S.: AfriSenti: A twitter sentiment analysis benchmark for African Languages. EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings pp. 13968–13981 (2 2023). https://doi.org/10.18653/v1/2023.emnlp-main.862, https://arxiv.org/abs/2302.08956v5
13. Niyongabo, R.A., Hong, Q., Kreutzer, J., Huang, L.: KINNEWS and KIRNEWS: Benchmarking cross-lingual text classification for Kinyarwanda and Kirundi. In: Scott, D., Bel, N., Zong, C. (eds.) Proceedings of the 28th International Conference on Computational Linguistics. pp. 5507–5521. International Committee on Computational Linguistics, Barcelona, Spain (Online) (Dec 2020). https://doi.org/10.18653/v1/2020.coling-main.480, https://aclanthology.org/2020.coling-main.480/
14. Ogueji, K., Zhu, Y., Lin, J., Cheriton, D.R.: Small data? no problem! exploring the viability of pretrained multilingual language models for low-resource languages pp. 116–126 (2021), https://github.com/google-research/bert/
15. Oladipo, A., Ogundepo, O., Ogueji, K., Lin, J., Cheriton, D.R.: An exploration of vocabulary size and transfer effects in multilingual language models for African languages, https://en.wikipedia.org/wiki/Ge'ez_script
16. Parvess, J.: BantuBERTa: Using language family grouping in multilingual language modeling for bantu languages (2023), https://repository.up.ac.za/handle/2263/92766, available at https://repository.up.ac.za/handle/2263/92766
17. Parvess, J., Marivate, V., Akinyi, V.: BantuBERTa model (2024). https://doi.org/10.57967/hf/3067, available at https://huggingface.co/dsfsi/BantuBERTa
18. Pham, T., Le, K.M., Tuan, L.A., Chi, H.: UniBridge: A unified approach to cross-lingual transfer learning for low-resource languages (6 2024), https://arxiv.org/abs/2406.09717v3
19. Ragni, A., Knill, K., Rath, S., Gales, M.: Data augmentation for low resource languages (9 2014), https://www.isca-speech.org/archive/interspeech_2014/i14_0810.html
20. Rasooli, M.S., Farra, N., Radeva, A., Yu, T., Mckeown, K.: Cross-lingual sentiment transfer with limited resources. Machine Translation **32**(1–2), 143–165 (Jun 2018). https://doi.org/10.1007/s10590-017-9202-6, https://doi.org/10.1007/s10590-017-9202-6
21. Savant, R., Shelke, A., Todmal, S., Kanphade, S., Joshi, A., Josh, R.: Universal cross-lingual text classification. In: 2024 IEEE 9th International Conference for Convergence in Technology (I2CT). pp. 1–6 (2024). https://doi.org/10.1109/I2CT61223.2024.10543381