What Happens When Small Is Made Smaller?

Busayo Awobade   Mardiyyah Oduwole   Steven Kolawole

MLCollective

Masakhane

Introductions

Over the years, large language models (LLMs) have exponentially increased in data and computational complexity. To ensure effective performance, most LLMs rely on abundant unlabeled text corpora, which unfortunately are limited for African languages [6]. Consequently, the absence or limited amount of African languages during the pre-training phase of LLMs leads to poor performance in these languages, presenting a predicament for natural language processing (NLP) tasks [5]. This scenario has been termed the "low-resource double-bind" by [2] to capture the coexistence of data and computation limitations on resources. Despite its prevalence in the NLP setting for low-resource languages, there is a notable lack of comprehensive research on the performance trade-offs involved.

A noteworthy initiative to address the scarcity of low-resource African languages in LLMs is the development of AfriBERTa, a language model based on XLM-RoBERTa specifically tailored to these languages [7]. AfriBERTa stands out as the first multilingual language model trained entirely from scratch on African languages, utilizing less than 1GB of data. Surpassing large language models like mBERT [4] and XLM-R [3], AfriBERTa has demonstrated exceptional performance in text categorization and named entity recognition (NER) tasks, producing competitive results. Remarkably, instead of relying on high-resource languages for transfer learning, AfriBERTa leverages linguistic similarities among low-resource languages, leading to promising outcomes. This approach proves highly advantageous for these languages and significantly impacts the sustainability of language models trained on limited datasets.

In this work, we investigate the effects of pruning, knowledge distillation, and quantization on AfriBERTa, a small-data language model exclusively trained on low-resource languages. Through a comprehensive series of experiments, we evaluate the effects of compression on model size, inference time, and performance across various metrics beyond accuracy.

Objectives

1. How tiny can we construct a small-data model using the knowledge distillation framework?
2. What are the efficiency, performance, and generalization limits of pruning on a small-data model?
3. What are the optimal reductions we can achieve in model size and inference time utilizing quantization methods without sacrificing accuracy?

Experimental Setup

For our experiments, we separately apply three compression techniques to the large and base AfriBERTa models [7].

- Compression Techniques
  - Pruning; unstructured magnitude pruning
  - Distillation; Task-specific and Task agnostic
  - Quantization; Dynamic Quantization and 8-bit Matrix Multiplication(LLM.int8())
- Datasets
  - AfriBERTa Corpus [7]; Was used for distilling knowledge from teacher models
  - MasakhaNER [1]; Used for downstream evaluations.
- **Models**; AfriBERTa base model and AfriBERTa large model.

Results for all experiments are average over three runs with different seeds.

Findings

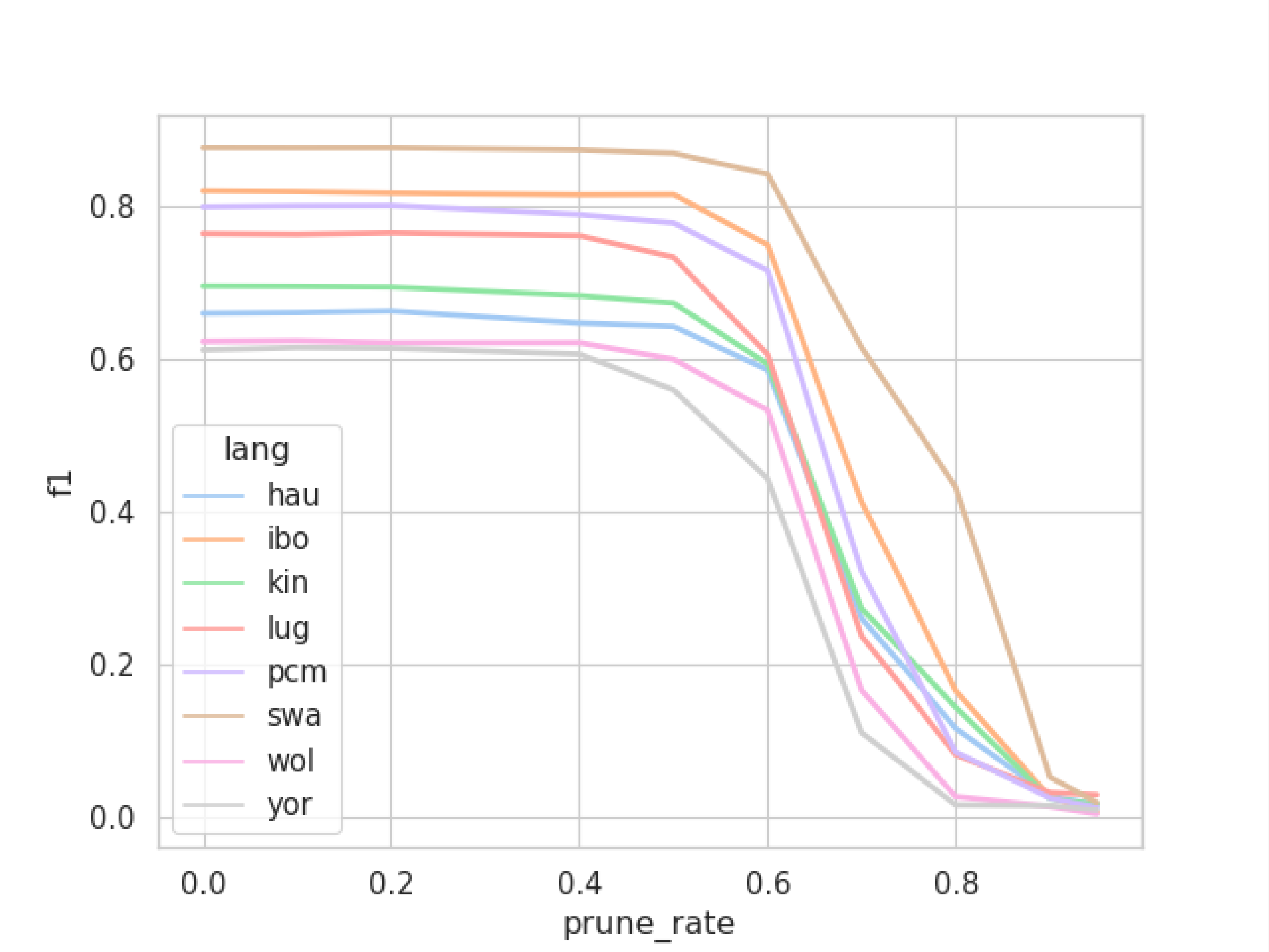

Figure 1. Mean F1 scores of languages over sparsity levels.

The limits of pruning for small-data pre-trained Models? We find that at 50% and 60% sparsity, the model maintained consistent performance. Some languages also showed moderate performance at 90% and 95% sparsity, indicating the model's potential robustness to pruning.

However, at 90% and 95% sparsity, certain languages like Yoruba and Luganda experienced significant performance reductions, possibly due to their specific linguistic characteristics and sparse datasets, making them more susceptible to pruning-induced degeneration.

To validate the findings, cross-lingual transfer experiments were conducted on AfriBERTa after pruning. The study highlights the importance of considering each language's unique qualities and dataset when determining the ideal sparsity level for pruning to achieve aggressive pruning while maintaining performance.

How does Pruning affect Out-of-Domain Generalization for this kind of model? The findings show that pruning can have a positive effect on OOD generalization for some languages, but its benefits are limited for others. Surprisingly, for many languages, including Swahili, pruned models perform as well as or even better than the original dense model up to 60% sparsity, achieving between 85% and 90% F1 score for Swahili.

However, for languages like Yoruba, which have higher linguistic complexity, even the dense model's performance is relatively low, with an F1 score of around 60%. The pruned models of such languages experience a significant drop in performance beyond 50% sparsity, revealing the challenge of compressing models with intricate linguistic structures.

Distillation strategy	Teacher	#Layers	#Att. Heads	#Params	amh	hau	ibo	kin	lug	luo	pcm	swa	wol	yor	avg
Task-agnostic	AfriBERTa-base	4	4	83M	64.96	87.28	83.58	68.15	74.57	63.02	78.92	83.89	55.46	73.62	73.35
		4	6	83M	64.23	87.34	83.84	67.59	74.60	60.00	79.40	84.00	57.21	73.38	73.16
		6	4	97M	65.96	87.60	85.55	70.16	75.90	64.61	81.65	85.48	57.70	74.82	74.94
		6	6	97M	66.92	87.91	85.28	69.81	77.19	68.40	81.72	85.08	60.28	75.47	**75.81**
	AfriBERTa-large	4	4	83M	65.28	87.28	84.15	68.83	73.82	63.79	79.80	84.13	56.30	73.43	73.68
		4	6	83M	65.25	87.62	84.28	68.82	74.66	62.60	78.88	84.07	55.11	73.85	73.51
		6	4	97M	69.38	88.25	85.08	69.49	75.44	63.87	82.71	85.87	56.42	73.89	75.04
		6	6	97M	71.98	88.72	85.76	71.76	78.30	68.10	84.24	87.07	61.45	77.61	77.50
Task-specific	AfriBERTa-base	4	4	83M	65.04	87.12	83.13	67.62	75.01	62.90	78.19	83.96	54.04	69.22	72.62
		4	6	83M	65.52	87.27	83.93	68.18	75.56	63.78	79.03	83.70	57.46	69.98	73.44
		6	4	97M	67.28	87.27	85.72	71.68	77.18	66.58	81.85	84.92	60.20	74.78	**75.75**
		6	6	97M	69.45	88.23	85.47	69.88	74.90	65.79	82.27	85.36	59.12	76.25	75.67
	AfriBERTa-large	4	4	83M	65.66	87.60	83.42	67.38	74.28	62.37	79.62	83.78	55.72	72.89	73.27
		4	6	83M	71.20	88.27	84.66	70.70	77.11	65.58	82.09	86.06	58.00	76.21	75.99
		6	4	97M	69.38	88.25	85.08	69.49	75.44	63.87	82.71	85.87	56.42	73.89	75.04
		6	6	97M	72.58	88.33	86.05	71.16	78.56	69.87	84.03	86.32	61.49	76.66	77.51

Table 1. **Results for the distilled models on NER task across 10 languages**. The scores are averaged over 3 runs using different seeds. The best student variant for each teacher is highlighted, and the best variant for each strategy is underlined.

How small can we make these models? We investigate this with two variations of distillation (task agnostic and task-specific) with the intent of maintaining competitive performance. Considering the ablation study in [7], we achieved up to 31% compression with only a 7% performance drop for the least-performing student model and a 1.9% decline compared to the best-performing original AfriBERTa model at 22% compression. The comparison between teacher and student models shows very competitive scores, with minimal differences for some languages. Interestingly, the student model trained by the large teacher even outperformed the base teacher on certain languages. Additionally, the task-agnostic models performed better in terms of F1 score compared to the task-specific models, with relatively minor differences.

Who is the best teacher? We examine the effectiveness of both base and large models at teaching the student models. The AfriBERTa large model produced the best-performing student but had a small 1.9% performance decline compared to the original AfriBERTa large model. However, the best-performing student by the base model only showed a 1.3% performance drop compared to the original base teacher, indicating that the base model is relatively better at transferring its knowledge to its students. As the attention head and layer ratio reduced, students taught using the base model caught up to those using the large model, with no noticeable difference in performance. These findings suggest that, when condensing knowledge into a compact student model with fewer parameters, the base model might be a more effective instructor. The instructor model selection significantly influences the performance of student models.

Language	Baseline	Dynamic	LLM.int8()
amh	26.01	12.78	13.27
hau	31.08	19.99	13.31
ibo	31.84	21.67	15.03
kin	27.19	20.95	16.85
lug	21.10	12.35	10.62
luo	22.40	5.53	5.47
pcm	41.70	17.96	16.34
swa	35.50	20.14	17.37
wol	25.38	20.95	14.78
yor	34.45	23.14	18.36

Inference Time Comparison (ms) for the Different Quantization Methods

Language	Baseline	Dynamic	LLM.int8()
Amh	**73.36**	68.02	73.28
Hau	89.93	85.35	**89.95**
Ibo	**86.96**	82.21	86.88
Kin	**73.98**	61.58	73.91
Lug	79.78	68.94	**79.83**
Luo	**70.04**	42.40	69.77
Pcm	**85.23**	74.37	85.18
Swa	87.89	84.58	**87.93**
Wol	**61.73**	47.36	61.71
Yor	**80.76**	65.10	80.74

Inference Time Comparison (ms) for the Different Quantization Methods

The effects of quantization on model performance and inference speed? We see the performance of quantized models vary across languages, with some outperforming the original dense model while others perform worse with negligible differences. The LLM.int8() shows superiority over dynamic quantization in terms of F1 scores, loss, and inference time across all languages, it achieves a 64.08% model size reduction and 52.3% inference time reduction. Quantization significantly reduces inference time for all languages, making it an effective technique for optimization even for small data pre-trained models for deployment on resource-limited devices. However, there is no one-size-fits-all solution, as quantization performance depends on factors like language, data type, and the quantization technique used.

Conclusion

We examined the impact of pruning, knowledge distillation, and quantization on the small-data language model, AfriBERTa, for low-resource languages. Our findings provide support to other findings relating to some of the works of compression on larger models as we extend those experiments to small-data pre-trained language models and find that these compression techniques also effectively improve the efficiency and performance of small-data models while maintaining a balance between efficiency, accuracy, and generalization capabilities.

Acknowledgement

We thank Nahid Alam and Naomi Saphra for assisting us with helpful suggestions and feedback and also ML Collective for the generous computational support.

Limitations

Some of the limitations of our work are;

- We solely focus on NER as the NLP task and no introduction of a new technique
- lacks coverage of language families beyond the selected 10 African languages, etc.

References

[1] D. Adelani, Jade Abbott, Graham Neubig, Daniel D'Souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, S. Muhammad, Chris C. Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, J. Alabi, Seid Muhie Yimam, Tajuddeen R. Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, V. Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin P. Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, C. Chukwuneke, N. Odu, Eric Peter Wairagala, S. Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane Mboup, D. Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima Diop, A. Diallo, Adewale Akinfaderin, T. Marengereke, and Salomey Osei. Masakhaner: Named entity recognition for african languages. *ArXiv*, abs/2103.11811, 2021.

[2] Orevaoghene Ahia, Julia Kreutzer, and Sara Hooker. The low-resource double bind: An empirical study of pruning for low-resource machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3316–3333, 2021.

[3] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, 2020.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational