AfroXLMR-Comet: Multilingual Knowledge Distillation with Attention Matching for Low-Resource Languages

Joshua Sakthivel Raju[1 [0009−0005−4414−2582]], Sanjay Somasundaram[1 [0009−0007−4172−1815]], Jaskaran Singh Walia[1 [0000−0002−9255−5446]], Srinivas R[1 [0009−0000−3669−4727]], and Vukosi Marivate[2,3 [0000−0002−6731−6267]]

[1] School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India
joshua.raju2604@gmail.com
[2] Data Science for Social Impact & African Institute for Data Science and Artificial Intelligence, University of Pretoria, South Africa
[3] Lelapa AI

Abstract. Language model compression through knowledge distillation has emerged as a promising approach for deploying large language models in resource-constrained environments. However, existing methods often struggle to maintain performance when distilling multilingual models, especially for low-resource languages. In this paper, we present a novel hybrid distillation approach that combines traditional knowledge distillation with a simplified attention matching mechanism, specifically designed for multilingual contexts. Our method introduces an extremely compact student model architecture, significantly smaller than conventional multilingual models. We evaluate our approach on five African languages: Kinyarwanda, Swahili, Hausa, Igbo, and Yoruba. The distilled student model–*AfroXLMR-Comet*–successfully captures both the output distribution and internal attention patterns of a larger teacher model (AfroXLMR-Large) while reducing the model size by over 85%. Experimental results demonstrate that our hybrid approach achieves competitive performance compared to the teacher model, maintaining an accuracy within 85% of the original model's performance while requiring substantially fewer computational resources. Our work provides a practical framework for deploying efficient multilingual models in resource-constrained environments, particularly benefiting applications involving African languages.

Keywords: Knowledge distillation · Parameter-efficient-training · Data-efficient training · NLP in resource-constrained settings.

1 Introduction

Large language models (LLMs) have become a pillar of modern Natural Language Processing (NLP), achieving state-of-the-art results across various tasks

[6,11,10]. Their performance only continues to improve with the expansion of computational power, the availability of vast datasets, and the scaling of model architectures [8,3,18]. But, despite their success, a significant portion of the world's languages, particularly low-resource languages (LRLs), remain underrepresented in NLP research. These languages, numbering in the thousands, lack the necessary linguistic resources and data required for traditional statistical methods to be effectively applied [16,5,19].

The challenges associated with LRLs are substantial and multifaceted. These languages often suffer from a lack of computational tools, limited digital presence, and insufficient educational infrastructure, which hinders their incorporation into modern NLP systems. While advancements in NLP for LRLs present immense potential, especially in regions such as Africa and India, where over 2.5 billion people speak these languages, the barriers are high. Addressing these challenges not only promises economic and cultural benefits but also plays a vital role in preserving linguistic heritage and improving societal outcomes, such as aiding in emergency response and enhancing educational and cultural exchanges [19].

1.1 Knowledge Distillation

Knowledge distillation (KD) refers to the process where a smaller model learns from a larger one, transferring knowledge from a teacher model to a student model to enhance the student's performance. The central concept is that the student model emulates the teacher model, using the teacher's insights to achieve competitive or even superior results. A typical KD framework includes three main components: knowledge, a distillation algorithm, and the teacher-student model architecture. The key challenge is to effectively convey knowledge from the large teacher model to the smaller student model while ensuring the student retains or improves its performance [4,2,7,20].

Response-based knowledge distillation focuses on aligning the final output (logits) of the student model with the teacher's logits, typically using soft targets and temperature scaling to capture the "dark knowledge" from the teacher. This approach has been widely applied in tasks like image classification, but it is limited by only utilizing the output of the last layer, potentially missing intermediate-level supervision that could benefit deeper networks [7,2].

To address this, feature-based knowledge distillation extends the method by transferring feature maps from intermediate layers of the teacher model to the student model, enhancing representation learning. Techniques like attention maps and activation boundaries are used to match features between layers, but challenges persist in selecting the right layers and handling size differences between layers [15,24].

In this work, we propose a novel hybrid distillation framework that integrates knowledge distillation and attention matching to improve multilingual model compression. Existing approaches often focus on either response-based distillation, which transfers knowledge through soft output distributions, or feature-based distillation, which aligns internal representations. Our method integrates

both, enabling a more comprehensive transfer of knowledge from teacher to student models. Additionally, we introduce a highly compact multilingual student model with a significantly smaller hidden dimension, optimized for low-resource African languages. Our contributions can be summarized as follows:

- **Hybrid Distillation Framework:** We propose a novel distillation approach that integrates knowledge distillation with attention matching, enabling the student model to learn both the output distribution and internal attention patterns of the teacher model.
- **Compact Multilingual Model:** We introduce an extremely compact multilingual architecture with a hidden dimension of 256, significantly smaller than existing models (typically 768 or higher), while maintaining reasonable performance.
- **Simplified Attention Matching:** Our mean-pooled attention matching mechanism effectively transfers knowledge between teacher and student models while reducing computational overhead compared to complex relation-based methods.
- **Evaluation on African Languages:** We conduct a systematic evaluation of our hybrid distillation approach on five African languages—Kinyarwanda (`rw`), Swahili (`sw`), Hausa (`ha`), Igbo (`ig`), and Yoruba (`yo`)—addressing gaps in model compression research for low-resource languages.
- **Empirical Analysis:** We analyze trade-offs between model size, computational efficiency, and performance, demonstrating that effective knowledge transfer is possible despite substantial architectural differences between teacher and student models.

2 Related Work

[1] proposed AfroXLMR, a multilingual pre-trained language model specifically adapted for African languages through multilingual adaptive fine-tuning (MAFT). Their approach has shown to enhance the performance of pre-trained models, such as XLM-R and AfriBERTa, on a variety of tasks for African languages by fine-tuning on monolingual texts from 17 high-resource African languages, alongside three widely spoken high-resource languages in Africa. One of the key innovations is the reduction of model size by removing tokens for non-African writing scripts from the embedding layer, which decreases the model size by approximately 50%. The resulting model not only performs competitively with language-adaptive fine-tuning (LAFT) on individual languages but also improves the cross-lingual transfer [17] abilities of models like XLM-R, while requiring significantly less disk space. This makes the AfroXLMR approach more efficient for practical deployment on tasks such as Named Entity Recognition (NER), topic classification, and sentiment analysis, especially in low-resource African languages.

In MiniLM [22], deep self-attention distillation is used to compress pre-trained Transformers by transferring knowledge from the teacher model to the

student model. MiniLMv2 [21] builds on this by introducing multi-head self-attention relation distillation for task-agnostic compression, where attention relations are defined as the scaled dot-product between query, key, and value pairs within the self-attention module. Unlike previous methods, MiniLMv2 allows the student model to have a different number of attention heads than the teacher, which removes the constraint of matching attention head numbers. By concatenating queries from multiple attention heads and splitting them to match the desired number of relation heads, MiniLMv2 enables more fine-grained attention knowledge transfer, leading to a deeper mimicry of the teacher's attention mechanisms. Moreover, MiniLMv2 examines layer selection beyond just the last layer, and finds that transferring knowledge from an upper-middle layer results in improved performance, especially for large models. Experimental results demonstrate that MiniLMv2, applied to both monolingual and multilingual pre-trained models, outperforms state-of-the-art methods, achieving better performance with fewer training examples and faster execution times.

3 Methodology

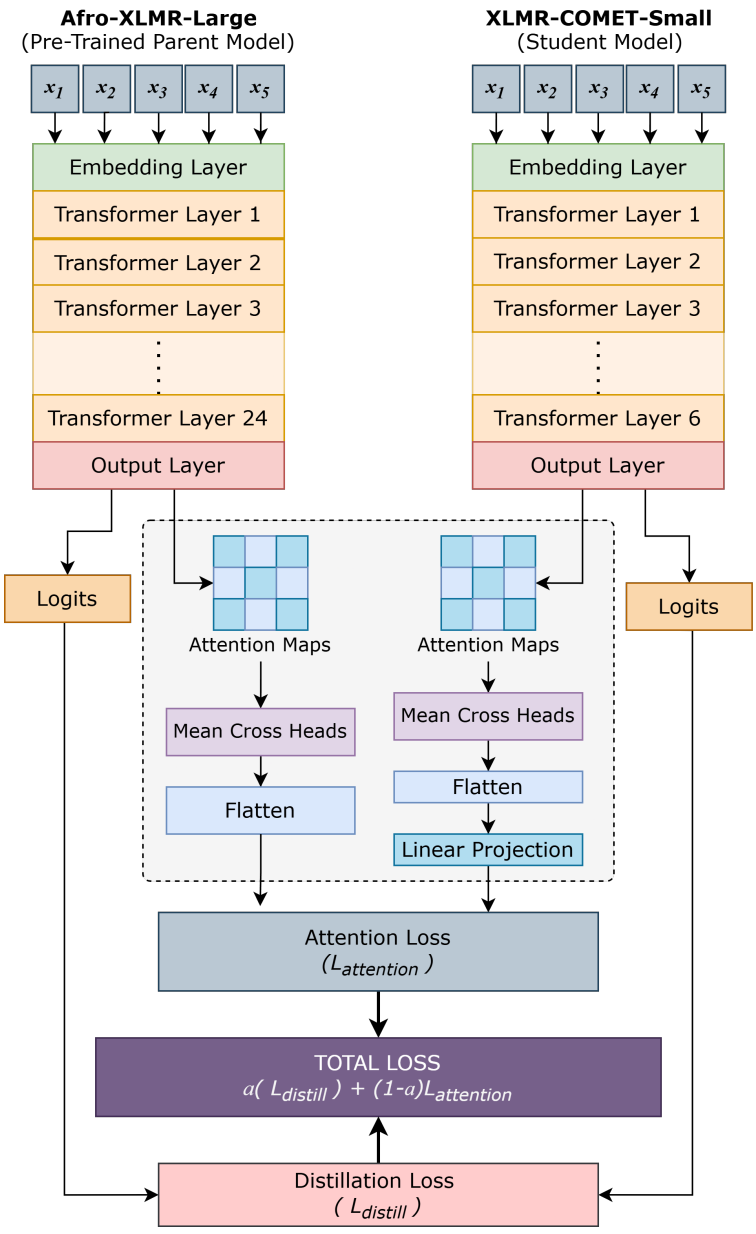

Fig. 1. Proposed hybrid distillation framework.

3.1 Model Architecture

Teacher Model For our teacher model, we utilize the Afro-XLM-RoBERTa-Large (AfroXLMR-Large) architecture, a multilingual language model specifically pre-trained on 17 African languages (Afrikaans, Amharic, Hausa, Igbo, Malagasy, Chichewa, Oromo, Naija, Kinyarwanda, Kirundi, Shona, Somali, Sesotho, Swahili, isiXhosa, Yoruba, and isiZulu) covering the major African language families and 3 high-resource languages (Arabic, French, and English). This model follows the standard transformer architecture with a hidden size of 1024, 16 attention heads, and 24 transformer layers (refer to Table 1). The model was pre-trained on a diverse corpus of African language texts, making it particularly suitable for our target languages.

Student Model Our student model architecture is derived from XLM-RoBERTa-COMET-small, leveraging the mMiniLM-L12xH384 XLM-R model proposed by [21]. However, we introduce significant modifications to further reduce the model size while maintaining performance (refer to Table 1). The key architectural changes include:

1. **Hidden Size Reduction:** We reduce the hidden size from the original 384 to 256 dimensions, to further decreasing the model's parameter count.

2. **Intermediate Layer:** To maintain architectural balance with the reduced hidden size, we adjust the intermediate layer size to 1024, compared to the original 1536 dimensions.

3. **Attention Head:** We configure the model with 8 attention heads, ensuring that each head operates on 32-dimensional key, query, and value vectors ($256/8 = 32$), maintaining the standard practice of having head dimensions that are factors of the hidden size.

This architectural configuration results in a substantially more compact model while preserving the essential multi-head attention mechanism necessary for capturing complex linguistic patterns. The reduced dimensionality is compensated for through our attention matching mechanism, which enables the student model to learn effective representations despite its smaller size.

Table 1. Comparison of model configurations and parameter counts, highlighting the reduction in parameters for the Student model.

Attribute	**AfroXLMR-Large**	**XLMR-Comet-Small**	**AfroXLMR-Comet**
Hidden Size	1024	384	256
Attention Heads	16	12	8
Hidden Layers	24	6	6
Intermediate Size	4096	1536	1024
Parameter Count	559,890,432	106,993,920	68,937,216

3.2 Dataset Preparation

In this work, we focus on a multilingual dataset–the MADLAD-400 dataset [9], a manually audited dataset based on CommonCrawl, spanning 419 languages. We employ a multilingual subset of the dataset that represents African languages, specifically Kinyarwanda (rw), Swahili (sw), Hausa (ha), Igbo (ig), and Yoruba (yo). The dataset is then split into 80% for training and 20% for validation. All sequences are tokenized to a maximum length of 128 tokens.

Table 2. Total and used sequences from the Madlad-400 dataset for selected languages

Lang.	**Sequences**	**Sequences Used**
Kinyarwanda	226,466	226,466
Swahili	537,847	250,000
Hausa	173,485	173,485
Igbo	54,410	54,410
Yoruba	52,067	52,067

3.3 Attention Matching Mechanism

Our work implements a simplified attention matching approach that focuses on the aggregate attention patterns rather than individual head-level interactions. While previous work, such as MiniLMv2 [21], addresses head count mismatches through relation heads and KL-divergence, we propose a more streamlined approach that captures the overall attention behavior.

We extract the attention matrices from the final layer of both teacher and student models. To manage the dimensional differences between the models, we first compute the mean across attention heads, producing a single attention map per sequence. This averaging operation reduces the attention tensors from shape *[batch_ size, num_ heads, sequence_ length, sequence_ length]* to *[batch_ size, sequence_ length, sequence_ length]*, effectively capturing the aggregate attention patterns across all heads.

The attention matrices are then flattened to vectors of shape *[batch_ size, sequence_ length* x *sequence_ length]*. To address the remaining dimensional differences between teacher and student models, we employ a learnable linear projection layer that maps the flattened student attention patterns to the teacher's dimensional space. The projection operation can be formalized as:

$$`A_{\text{student\_proj}} = W \cdot A_{\text{student\_flat}} \tag{1}$$

where:

- $A_{\text{student\_flat}}$ is the flattened student attention matrix.
- W is the learned projection matrix.
- $A_{\text{student\_proj}}$ is the projected student attention pattern.

Algorithm 1 Multilingual Knowledge Distillation

Require: Student model S, Teacher model T, Temperature τ, Weight factor α, Projection layer P
Require: Training dataset $\mathcal{D}$, Loss function $\mathcal{L}_{\text{MSE}}$
for each batch $(X, Y) \in \mathcal{D}$ **do**
  Forward Pass:
  $Z_T \leftarrow T(X)$ ▷ Teacher logits (no gradient)
  $Z_S \leftarrow S(X)$ ▷ Student logits
  Distillation Loss:
  $P_T \leftarrow \text{softmax}(Z_T/\tau)$
  $P_S \leftarrow \text{softmax}(Z_S/\tau)$
  $\mathcal{L}_{\text{distill}} \leftarrow -\sum P_T \log(P_S + \epsilon)$
  if attention available in S and T **then**
    Attention Loss:
    $A_T \leftarrow \text{mean}(T.\text{attentions}[-1], \text{dim} = 1)$
    $A_S \leftarrow \text{mean}(S.\text{attentions}[-1], \text{dim} = 1)$
    $A_T \leftarrow \text{flatten}(A_T), \quad A_S \leftarrow \text{flatten}(A_S)$
    $A'_S \leftarrow P(A_S)$ ▷ Project student attention
    $\mathcal{L}_{\text{attn}} \leftarrow \mathcal{L}_{\text{MSE}}(A'_S, A_T)$
    $\mathcal{L} \leftarrow \alpha\mathcal{L}_{\text{distill}} + (1 - \alpha)\mathcal{L}_{\text{attn}}$
  else
    $\mathcal{L} \leftarrow \mathcal{L}_{\text{distill}}$
  end if
  Backward Pass:
  Update S using $\nabla\mathcal{L}$
end for

The training objective for attention matching is computed using Mean Squared Error (MSE) loss between the projected student attention and the flattened teacher attention:

$$L_{\text{attention}} = \text{MSE}(A_{\text{student\_proj}}, A_{\text{teacher\_flat}}) \tag{2}$$

where:

- $L_{\text{attention}}$ is the attention loss function.
- MSE is the Mean Squared Error function.
- $A_{\text{student\_proj}}$ is the projected student attention pattern.
- $A_{\text{teacher\_flat}}$ is the flattened teacher attention matrix.

This approach, while simpler than the relation-based methods, proves effective in practice. By focusing on the aggregate attention patterns rather than individual head interactions, we reduce computational complexity while still capturing the essential aspects of the teacher's attention mechanism. The learned projection layer allows the student to adapt its attention patterns to match the teacher's higher-dimensional space, facilitating effective knowledge transfer despite the architectural differences between the models.

Our final loss function combines this attention matching loss with traditional knowledge distillation:

$$L_{\text{total}} = \alpha \cdot L_{\text{distill}} + (1 - \alpha) \cdot L_{\text{attention}} \tag{3}$$

where:

- $\alpha = 0.5$ balances between the distillation and attention matching objectives.

3.4 Training Process

To facilitate effective knowledge transfer between these architecturally different models, we employ a two-stage training process (refer Figure 1). First, we initialize the student model with the reduced configuration parameters. Then, we apply our combined distillation approach, using both soft target probabilities and attention matching to transfer knowledge from the teacher to the student model.

Instead of task-specific distillation, we follow a task-agnostic knowledge distillation framework. The student model is trained on a general language modeling objective, learning to replicate the teacher's predictions and attention patterns on unlabeled text data. By distilling general linguistic representations through this pre-training task, rather than optimizing for specific downstream applications, our model remains adaptable to various NLP tasks without requiring task-specific retraining.

For soft target distillation, we use a temperature parameter T=2.0 to create smoothed probability distributions from both models' logits, then measure their difference using Kullback-Leibler (KL) divergence. This smoothing reveals the teacher's underlying knowledge through relative probabilities across all classes, while KL divergence effectively captures how the student's probability distribution diverges from the teacher's desired distribution. The soft probability for class i is calculated as:

$$p_i = \frac{\exp(z_i/T)}{\sum_{j=1}^{N} \exp(z_j/T)} \tag{4}$$

where:

- z_i is the logit (raw score) for class i
- T is the temperature parameter (typically T>1)
- N is the total number of classes
- p_i is the resulting soft probability for class i

The training process is optimized using AdamW optimizer with a learning rate of *5e-5*. We implement early stopping with a patience of 3 epochs and a threshold of 0.01 to prevent overfitting, ensuring the student model achieves optimal performance without unnecessary computation. Refer to Table 3 for other training parameters for the model distillation process, Figure 2 for the training and validation loss curves, and to Algorithm 1 for the training process' algorithm.

Table 3. Training parameters and values used for model distillation.

Parameter	Value
Max Sequence Length	128
Padding	True
Truncation	True
Evaluation Strategy	Epoch
Logging Steps	50
Learning Rate	5×10^{-5}
Number of Training Epochs	15
Per Device Train Batch Size	8
Gradient Accumulation Steps	2
Mixed Precision (FP16)	True
Save Strategy	Epoch
Save Total Limit	3
Load Best Model at End	True
Early Stopping Patience	3
Early Stopping Threshold	0.01

4 Experiments

We evaluate our distilled student model–AfroXLMR-Comet–using the AfriSenti dataset from the AfriSenti-SemEval [12,13,14,23] Shared Task 12, which covers 17 African languages: Hausa, Yoruba, Igbo, Nigerian Pidgin, Amharic, Tigrinya, Oromo, Swahili, Algerian Arabic, Kinyarwanda, Twi, Mozambican Portuguese, Moroccan Arabic, Fongbe, Lingala, Kamba, and Luganda. For our comparative analysis with AfroXLMR models, we focus on five major languages: *Kinyarwanda, Swahili, Hausa, Igbo,* and *Yoruba*. The dataset consists of manually annotated tweets categorized into positive, negative, and neutral sentiments (refer Table 4). We maintain the original data splits provided by the task organizers for consistent comparison with benchmark models (refer Table 5). Each model was fine-tuned for 5 epochs at a learning rate of *2e-5* on each language dataset provided by the task and then retrained for Task A. Our benchmark results are presented in Table 6.

Table 4. Field descriptions of the dataset.

Field	Description
ID	Alpha-Numeric Serial Numbers
Tweet	Tweet Content
Label	Tweet Sentiment Label

Further in Table 7, the parameter count, inference time, and model size comparison highlight the efficiency of the proposed AfroXLMR-Comet model. AfroXLMR-Large, with the highest parameter count (559.9M), also has the largest model size (2.09 GB) and the slowest inference time (293.9 milliseconds)

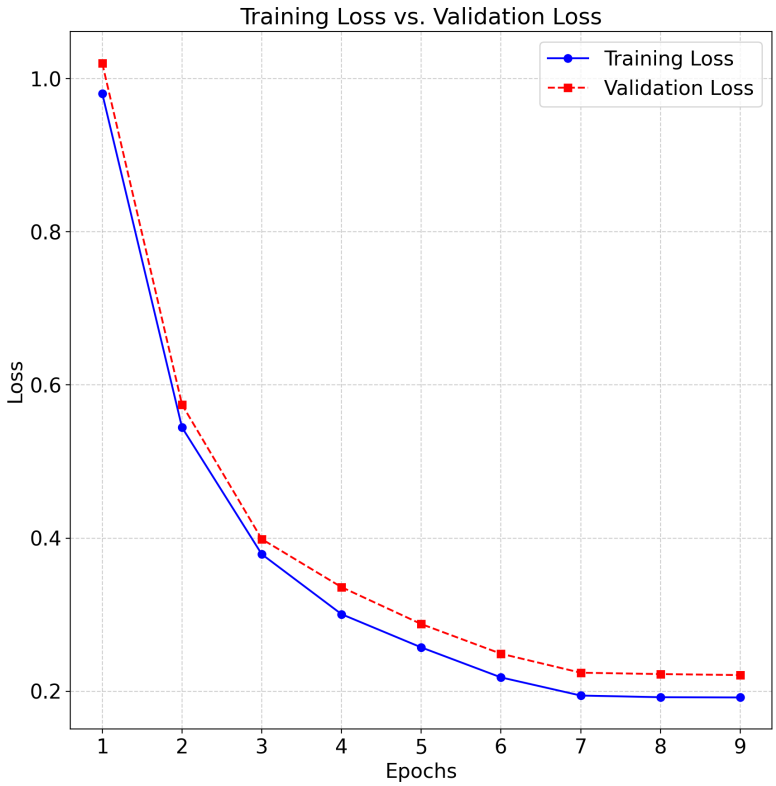

Fig. 2. Training and validation loss curves showing model convergence over 9 epochs. Training terminated via early stopping to prevent overfitting.

Table 5. Number of training and testing samples for each language in the AfriSenti dataset

Language	**Train**	**Test**
Kinyarwanda	3,302	1,026
Swahili	1,810	748
Hausa	14,172	5,303
Igbo	10,192	3,682
Yoruba	8,522	4,515

due to its significant computational requirements. AfroXLMR-Base (278M parameters, 1.04 GB) provides a better balance, achieving an inference time of 100.5 milliseconds, while AfroXLMR-Mini (117.6M parameters, 448.79 MB) further reduces computational load with a faster inference time of 30.2 milliseconds. Our AfroXLMR-Comet model, with just 68.9M parameters and a compact size of 262.99 MB, achieves the fastest inference time of 14.0 milliseconds, making it the most efficient among all models.

Table 6. Comparison of sentiment classification models on the AfriSenti-SemEval dataset. The table presents F1 scores on test sets after 5 training epochs, with AfroXLMR-Large as the parent model and AfroXLMR-Comet as the student. The last column indicates the average F1 score across all languages.

Model	kin	swa	hau	igo	yor	avg
AfroXLMR-Large	68.91	64.10	79.30	79.60	74.20	73.22
AfroXLMR-Base	62.54	60.81	78.27	78.63	70.25	70.10
AfroXLMR-Mini	60.23	61.88	74.29	75.20	63.10	66.94
AfroXLMR-Comet	58.94	55.30	71.89	73.82	66.39	65.28

Table 7. Comparison of AfroXLMR-Large, AfroXLMR-Base, AfroXLMR-Mini, and our distilled model in terms of parameter count, inference time, and model size. The proposed distilled model is significantly smaller and the fastest among the models.

Model	# Params	Inf. Time (ms)	Size (MB)
AfroXLMR-Large (Parent)	559,890,432	293.9	2135.86
AfroXLMR-Base	278,043,648	100.5	1060.67
AfroXLMR-Mini	117,640,704	30.2	448.79
AfroXLMR-Comet (Student)	68,937,216	14.0	262.99

5 Results and Discussion

The distilled student model–AfroXLMR-Comet–achieved a substantial parameter reduction of 87.69%, compressing the teacher model's 559,890,432 parameters down to 68,937,216. This reduction in size results in a trade-off in performance, as the student model achieves an average F1 score of 65.28% compared to Afro-XLMR-Large's 73.22% on the AfriSenti-SemEval sentiment classification benchmark. Despite this drop in accuracy, the student model significantly reduces inference latency and memory footprint, making it particularly suitable for deployment in resource-constrained environments. Additionally, the student model's compact size of 262.99 MB (compared to 2.09 GB for Afro-XLMR-Large) highlights its efficiency in handling low-resource languages.

Notably, the distilled student model's performance is very close to Afro-XLMR-Mini, which achieves an F1 score of 66.94% but is almost twice the size at 448.79 MB, further showcasing that the student model balances efficiency and performance, achieving comparable results with a significantly smaller memory footprint. The simplification of attention mechanisms further emphasizes the practicality of this approach, ensuring efficient knowledge transfer while maintaining competitive performance across multiple African languages.

6 Conclusion

In this work, we introduced a hybrid distillation framework for the efficient compression of pre-trained multilingual transformer models, specifically focus-

ing on African languages. By integrating task-agnostic knowledge distillation, attention matching, and adaptive learning mechanisms, we successfully distilled AfroXLMR-Large into a lightweight yet effective student model. Our approach achieves a significant reduction in computational and memory requirements while maintaining competitive performance, demonstrating a viable trade-off between efficiency and accuracy.

Looking ahead, future work will explore extending this methodology to other language families and investigating domain-specific fine-tuning to further enhance the adaptability of compressed multilingual models. Additionally, we aim to refine attention transfer mechanisms to improve performance retention in extreme compression settings. Our findings highlight the transformative potential of efficient model compression in democratizing access to NLP for low-resource languages, ensuring that cutting-edge language technologies become more accessible to linguistic communities worldwide.

7 Limitations

While our framework shows promising results, several important limitations should be acknowledged:

- **Projection Layer Computation:** The projection layer matching between teacher and student attention matrices introduces additional computational overhead. For a sequence length of 128, the projection layer requires maintaining and computing gradients for a large parameter matrix of size $(128^2) \times (128^2)$, which can be memory-intensive during training. This limitation becomes more pronounced when dealing with longer sequences or larger batch sizes, leading to memory allocation challenges during both fine-tuning and inference. This limitation, while less severe than head-level matching approaches, still impacts the model's deployability in resource-constrained environments.
- **Data and Resource Constraints:** A fundamental challenge in our work is the scarcity of high-quality dataset samples for many low-resource languages, which inherently limits the extent of multilingual generalization. Due to computational constraints, our experimental validation was limited to five languages.
- **Temperature Sensitivity:** The distillation process relies heavily on the temperature parameter τ in the softmax computations. While we use a fixed temperature of 2.0, the optimal temperature may vary across different languages and tasks. Our approach lacks adaptive temperature scaling that could potentially improve knowledge transfer for specific language combinations.
- **Limited Comparative Analysis:** Our evaluation does not include systematic comparisons with other state-of-the-art knowledge distillation methods or recent compressed multilingual models. While we demonstrate improvements over baseline models within our framework, a comprehensive benchmarking against competing approaches or other recent distillation techniques

for low-resource languages would provide stronger evidence of our method's relative performance. We acknowledge this as a significant gap and plan to conduct more extensive comparative evaluations in future work.

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.

Data Availability Statement. The trained AfroXLMR-Comet model is publicly available on the Hugging Face repository of the DSFSI Research Group: https://huggingface.co/dsfsi/afro-xlmr-comet.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alabi, J.O., Adelani, D.I., Mosbach, M., Klakow, D.: Adapting pre-trained language models to african languages via multilingual adaptive fine-tuning (2022), https://arxiv.org/abs/2204.06487, last accessed 2024/10/26
2. Ba, J., Caruana, R.: Do deep nets really need to be deep? In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K. (eds.) Advances in Neural Information Processing Systems. vol. 27. Curran Associates, Inc. (2014), https://proceedings.neurips.cc/paper_files/paper/2014/file/ea8fcd92d59581717e06eb187f10666d-Paper.pdf, last accessed 2024/10/25
3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 1877–1901. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf, last accessed 2024/10/26
4. Buciluǎ, C., Caruana, R., Niculescu-Mizil, A.: Model compression. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 535–541 (2006)
5. Cieri, C., Maxwell, M., Strassel, S., Tracey, J.: Selection criteria for low resource language programs. In: Calzolari, N., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 4543–4549. European Language Resources Association (ELRA), Portorož, Slovenia (May 2016), https://aclanthology.org/L16-1720/, last accessed 2024/10/25

6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2019), https://arxiv.org/abs/1810.04805, last accessed 2025/10/1
7. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network (2015), https://arxiv.org/abs/1503.02531, last accessed 2025/10/1
8. Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models (2020), https://arxiv.org/abs/2001.08361, last accessed 2024/10/26
9. Kudugunta, S., Caswell, I., Zhang, B., Garcia, X., Xin, D., Kusupati, A., Stella, R., Bapna, A., Firat, O.: Madlad-400: A multilingual and document-level large audited dataset. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 67284–67296. Curran Associates, Inc. (2023), https://proceedings.neurips.cc/paper_files/paper/2023/file/d49042a5d49818711c401d34172f9900-Paper-Datasets_and_Benchmarks.pdf, last accessed 2024/10/27
10. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension (2019), https://arxiv.org/abs/1910.13461, last accessed 2025/10/1
11. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach (2019), https://arxiv.org/abs/1907.11692, last accessed 2024/10/26
12. Muhammad, S., Abdulmumin, I., Ayele, A., Ousidhoum, N., Adelani, D., Yimam, S., Ahmad, I., Beloucif, M., Mohammad, S., Ruder, S., Hourrane, O., Jorge, A., Brazdil, P., Ali, F., David, D., Osei, S., Shehu-Bello, B., Lawan, F., Gwadabe, T., Rutunda, S., Belay, T., Messelle, W., Balcha, H., Chala, S., Gebremichael, H., Opoku, B., Arthur, S.: AfriSenti: A Twitter sentiment analysis benchmark for African languages. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 13968–13981. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.emnlp-main.862, https://aclanthology.org/2023.emnlp-main.862
13. Muhammad, S.H., Abdulmumin, I., Yimam, S.M., Adelani, D.I., Ahmad, I.S., Ousidhoum, N., Ayele, A.A., Mohammad, S.M., Beloucif, M., Ruder, S.: SemEval-2023 Task 12: Sentiment Analysis for African Languages (AfriSenti-SemEval). In: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023). Association for Computational Linguistics (2023)
14. Muhammad, S.H., Adelani, D.I., Ruder, S., Ahmad, I.S., Abdulmumin, I., Bello, B.S., Choudhury, M., Emezue, C.C., Abdullahi, S.S., Aremu, A., orge, A., Brazdil, P.: NaijaSenti: A Nigerian Twitter sentiment corpus for multilingual sentiment analysis. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. pp. 590–602. European Language Resources Association, Marseille, France (Jun 2022), https://aclanthology.org/2022.lrec-1.63, last accessed 2024/10/28
15. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets (2015), https://arxiv.org/abs/1412.6550, last accessed 2024/10/25

16. Singh, A.K.: Natural language processing for less privileged languages: Where do we come from? where are we going? In: Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages (2008)
17. Thangaraj, H., Chenat, A., Walia, J.S., Marivate, V.: Cross-lingual transfer of multilingual models on low resource african languages (2024), https://arxiv.org/abs/2409.10965, last accessed 2024/11/19
18. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models (2023), https://arxiv.org/abs/2302.13971, last accessed 2025/10/1
19. Tsvetkov, Y.: Opportunities and challenges in working with low-resource languages. Slides Part-1 p. 39 (2017), last accessed 2024/10/26
20. Urban, G., Geras, K.J., Kahou, S.E., Aslan, O., Wang, S., Caruana, R., Mohamed, A., Philipose, M., Richardson, M.: Do deep convolutional nets really need to be deep and convolutional? (2017), https://arxiv.org/abs/1603.05691, last accessed 2024/10/26
21. Wang, W., Bao, H., Huang, S., Dong, L., Wei, F.: Minilmv2: Multi-head self-attention relation distillation for compressing pretrained transformers (2021), https://arxiv.org/abs/2012.15828, last accessed 2025/10/1
22. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., Zhou, M.: Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers (2020), https://arxiv.org/abs/2002.10957, last accessed 2025/10/1
23. Yimam, S.M., Alemayehu, H.M., Ayele, A., Biemann, C.: Exploring Amharic sentiment analysis from social media texts: Building annotation tools and classification models. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 1048–1060. Barcelona, Spain (Online) (Dec 2020)
24. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer (2017), https://arxiv.org/abs/1612.03928, last accessed 2024/10/26