On Multilingual Encoder Language Model Compression for Low-Resource Languages

Daniil Gurgurov[1,2] **Michal Gregor**[3] **Josef van Genabith**[1,2] **Simon Ostermann**[1,2,4]
[1]Saarland University
[2]German Research Center for Artificial Intelligence (DFKI)
[3]Kempelen Institute of Intelligent Technologies (KInIT)
[4]Centre for European Research in Trusted AI (CERTAIN)
{daniil.gurgurov, josef.van_genabith, simon.ostermann}@dfki.de, michal.gregor@kinit.sk

Abstract

In this paper, we combine two-step knowledge distillation, structured pruning, truncation, and vocabulary trimming for extremely compressing multilingual encoder-only language models for low-resource languages. Our novel approach systematically combines existing techniques and takes them to the extreme, reducing layer depth, feed-forward hidden size, and intermediate layer embedding size to create significantly smaller monolingual models while retaining essential language-specific knowledge. **We achieve compression rates of up to 92% while maintaining competitive performance, with average drops of 2–10% for moderate compression and 8–13% at maximum compression** in four downstream tasks, including sentiment analysis, topic classification, named entity recognition, and part-of-speech tagging, across three low-resource languages. Notably, the performance degradation correlates with the amount of language-specific data in the teacher model, with larger datasets resulting in smaller performance losses. Additionally, we conduct ablation studies to identify the best practices for multilingual model compression using these techniques.

1 Introduction

Small multilingual encoder language models (LMs), such as mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), and Glot-500m (Imani et al., 2023), have demonstrated strong performance across a diverse range of low-resource languages (Hu et al., 2020; Asai et al., 2024), often outperforming large-scale proprietary models on various sequential tasks (Adelani et al., 2024; Gurgurov et al., 2025). However, even these relatively compact multilingual models may still be excessively large for use in individual languages due to redundant capacity and expensive inference (Singh and Lefever, 2022; Cruz, 2025).

To address this, we propose a novel combination of model compression approaches for transforming multilingual encoder-only models into maximally small, efficient, language-specific alternatives while retaining competitive performance. Our methodology integrates knowledge distillation (Hinton et al., 2015), structured pruning (Kim and Hassan, 2020; Hou et al., 2020), weight truncation, and vocabulary trimming (Abdaoui et al., 2020; Ushio et al., 2023) to systematically reduce model size by compressing the depth (number of layers), feed-forward intermediate width, hidden size, and tokenizer vocabulary. Our experiments demonstrate that this pipeline achieves compression rates of up to 92%, with performance drops of 2-10% for moderate compression (up to 87%) and 8-13% at maximum compression on downstream tasks such as sentiment analysis, topic classification, named entity recognition, and part-of-speech tagging. Notably, for moderate compression levels, the extent of degradation depends more on the strength of the teacher model than on the compression itself.

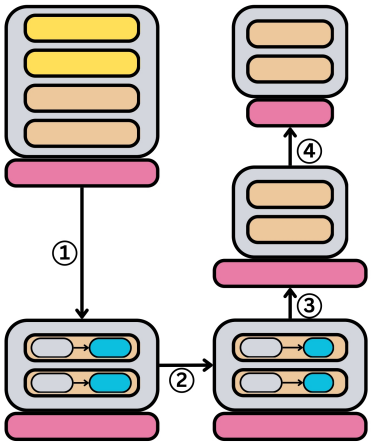


Figure 1: Overview of our multilingual model compression methodology. We use **(1)** knowledge distillation to reduce layers, **(2)** structured pruning to eliminate redundant feed-forward network width, and **(3)** hidden size reduction and another round of knowledge distillation from the previous student model. Finally, **(4)** vocabulary trimming is applied to retain language-specific tokens.

Proceedings of the IJCNLP-AACL 2025 Student Research Workshop, pages 47–58

Beyond compression, we investigate the impact of using multilingual versus monolingual teacher models, evaluate different initialization strategies for knowledge distillation, and analyze additional compression variables. Our findings contribute to the development of highly efficient, environmentally friendly models (Strubell et al., 2020) for low-resource languages and explore how strongly models can be compressed. The code for our experiments is made publicly available at https://github.com/d-gurgurov/Multilingual-LM-Disitillation.

2 Methodology

In this section, we present our multilingual model compression strategy, illustrated in Figure 1. Our approach combines several existing compression techniques in a novel way that, to the best of our knowledge, has not been explored in this combination within the multilingual context.

2.1 Layer Reduction via Knowledge Distillation

We reduce the number of transformer layers in the teacher model by half to obtain an initial compact student model (Sanh et al., 2020). The student is initialized with the layers of the teacher and trained using a combination of Masked Language Modeling (MLM) (Devlin et al., 2019) and Mean Squared Error (MSE) loss for knowledge distillation (Hinton et al., 2015) for 10 epochs. Both losses are weighted equally (α=0.5, though other values were explored; see Appendix 8). The teacher is a multilingual encoder fine-tuned on the target language (see Section 4).

2.2 Width Reduction via Structured Pruning

We apply structured pruning (Kim and Hassan, 2020) to reduce the intermediate size of the feed-forward layers from 3072 to 2048. Neuron importance is estimated using first-order gradient information accumulated from forward and backward passes over MLM validation data. At each layer, neurons are ranked by their absolute gradient values, and the least important ones are removed based on a target pruning ratio. The remaining neurons are then reordered to preserve model functionality. For consistency, the same pruning ratio is applied across all layers.

2.3 Hidden Size Compression with Secondary Knowledge Distillation

We compress the hidden embedding dimension from 768 to either 312, 456, or 564 via truncation, retaining the first k dimensions.[1] A second round of knowledge distillation is then performed, using the width-reduced model from the previous step as the new teacher, similar to Wang et al. (2023), with training for 10 epochs.

2.4 Vocabulary Reduction

We reduce the vocabulary size by selecting the top 40,000 most frequent tokens from a target-language corpus, along with their corresponding embeddings (Ushio et al., 2023). This ensures that the resulting model retains only language-specific tokens, which significantly reduces the overall model size.

3 Experiments

Below, we describe the datasets, languages, tasks, and baseline systems used in our evaluation.

3.1 Knowledge Distillation Data

We use GlotCC (Kargaran et al., 2025), a large-scale multilingual corpus derived mainly from CommonCrawl (Wenzek et al., 2020), as the primary dataset for both stages of knowledge distillation. Data distributions for the selected languages are reported in Appendix F. We use GlotCC for training, and the FLORES-200 development set (Team et al., 2022) for validation during training.

3.2 Languages and Tasks

We evaluate our models on four tasks: Topic Classification (TC), Sentiment Analysis (SA), Named Entity Recognition (NER), and Part-of-Speech Tagging (POS), covering three low-resource languages–Maltese, Slovak, and Swahili (Joshi et al., 2020). For TC, we use the 7-class SIB-200 dataset (Adelani et al., 2024), and for SA, we compile binary sentiment datasets from multiple sources (Dingli and Sant, 2016; Cortis and Davis, 2019; Pecar et al., 2019; Muhammad et al., 2023a,b). For NER, we use WikiANN (Pan et al., 2017), and for POS, we use Universal Dependencies v2.15 (de Marneffe et al., 2021) and MasakhaPOS (Dione et al., 2023). For all tasks, we train Sequential Bottleneck task adapters (Pfeiffer et al., 2020) with fixed hyperparameters (see Appendix H). Performance is mea-

[1] The hidden size must be divisible by the number of attention heads.

sured using macro-averaged F1 (Sokolova et al., 2006) for TC and SA, and "seqeval" F1 (Nakayama, 2018) for NER and POS.

3.3 Models and Baselines

We compress two encoder multilingual models–mBERT (Devlin et al., 2019) and XLM-R-base (Conneau et al., 2020)–adapted to target languages through fine-tuning on language-specific data, and compare the reduced models to two baselines: (1) the original, non-adapted models, and (2) language-adapted versions. In both cases, we train an identical task adapter using the same task-specific datasets as for the compressed models.

4 Findings

Our key findings are outlined below.

4.1 Distillation

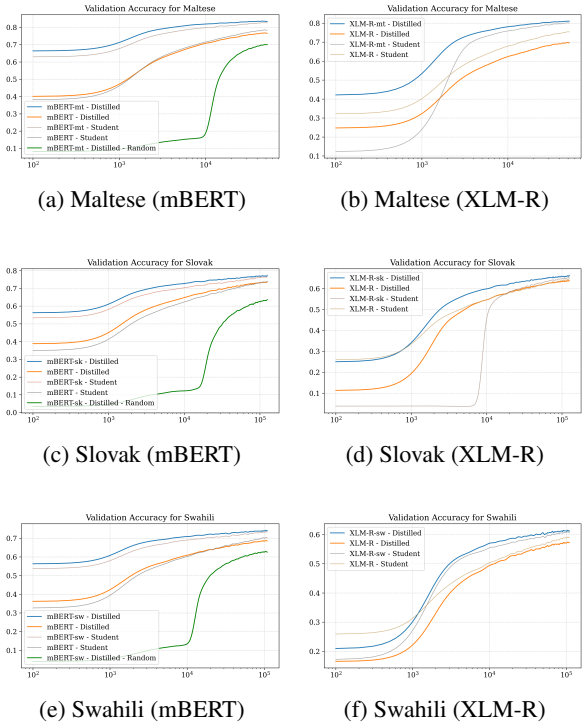
Distilling knowledge from a multilingual teacher into a monolingual student model is less effective than using a target-language adapted teacher, as evidenced by the differences in validation accuracies shown in Figure 2. This discrepancy possibly stems from the multilingual teacher's broad cross-lingual representations, which are not directly aligned with the requirements of a monolingual student. In contrast, monolingual teachers provide more targeted, language-specific representations, resulting in better student performance.

Distillation loss: We compare KL divergence and MSE as distillation loss functions, and observe that MSE leads to better and faster convergence (Appendix A), in line with prior work (Kim et al., 2021; Nityasya et al., 2022).

4.2 Weight Initialization

Weight initialization plays a crucial role in training the student model, with knowledge distillation providing only a marginal additional performance improvement (Figure 2). This partly aligns with the findings of Wibowo et al. (2024), who explored distilling multilingual abilities for multilingual tasks, whereas our focus is on monolingual distillation. Training a student-sized model initialized with teacher weights, but without knowledge distillation, results in a slight performance drop compared to a fully distilled model.

Initialization Strategies: Among various initialization strategies, initializing the student with the last k layers for mBERT and every other layer



(a) Maltese (mBERT) (b) Maltese (XLM-R)

(c) Slovak (mBERT) (d) Slovak (XLM-R)

(e) Swahili (mBERT) (f) Swahili (XLM-R)

Figure 2: First-step KD validation accuracies for mBERT and XLM-R with models initialized using the last k layers. mBERT- and XLM-R-mt, sk, sw refer to models adapted to the target language; *distilled* denotes models trained with distillation loss, while *student* refers to identically trained models without distillation loss. The best accuracy is in all cases achieved when distilling from a target-language adapted model.

(stride) for XLM-R consistently outperforms alternatives such as using the first k layers and combining first and last layers (Appendix B). Random initialization performs significantly worse, emphasizing the importance of weight reuse (Sun et al., 2019; Singh and Lefever, 2022).

4.3 Pruning and Truncation

Distilled models can be compressed further using structured pruning, hidden size reduction, and vocabulary trimming, while maintaining competitive performance.

Intermediate size reduction: Reducing the intermediate size of feed-forward layers from 3072 to 2048 via structured pruning results in negligible performance loss (Table 1). However, more aggressive reductions degrade quality significantly, making 2048 a practical lower bound. We do not prune attention heads, as removing even a minimal number (e.g., three) causes severe degradation (>50% performance drop in preliminary experiments).

Hidden size reduction: We reduce the hidden

Compression Stage	Params	Size	Task Performance (F1)												Avg
			Maltese				Slovak				Swahili				
			TC	SA	NER	POS	TC	SA	NER	POS	TC	SA	NER	POS	
Baselines															
Multilingual	279M	279M	68.1	56.0	54.3	89.9	88.1	95.6	91.1	97.3	78.4	81.5	84.6	89.4	81.2
Language-adapted	279M	279M	85.0	76.2	69.2	95.4	86.2	94.8	91.0	97.1	87.5	84.1	82.7	89.2	**86.5**
Compression Pipeline (minimal degradation)															
Layer reduction	236M (-15%)	236M	84.0	77.2	63.5	94.3	86.3	92.9	90.1	96.3	82.9	81.3	82.9	89.2	85.1
+ FFN pruning	226M (-20%)	226M	84.7	78.6	60.1	94.2	86.1	93.4	90.0	96.1	82.4	82.7	83.6	89.5	85.1
+ Hidden 564	163M (-40%)	163M	83.4	74.9	53.0	93.7	84.9	92.7	89.1	96.8	85.8	81.0	80.8	89.4	83.8
+ Vocabulary	45M (-85%)	45M	84.1	72.4	60.9	93.0	85.3	92.9	89.3	96.4	85.7	80.9	82.0	89.1	**84.3**
Further compression (moderate degradation)															
+ Hidden 456	131M (-53%)	131M	78.5	69.9	62.5	92.7	86.0	93.0	88.3	96.3	83.1	79.3	80.7	88.9	83.3
+ Vocabulary	35M (-87%)	35M	78.5	70.7	63.3	92.5	86.1	92.9	88.4	96.3	82.5	79.0	80.2	89.0	83.3
Maximum compression (higher degradation)															
+ Hidden 312	89M (-68%)	89M	66.9	70.1	35.7	87.6	84.0	90.9	88.0	95.5	76.4	80.1	80.7	88.3	78.7
+ Vocabulary	23M (-92%)	23M	67.2	71.4	37.1	87.5	84.0	90.5	88.2	95.6	78.0	80.5	79.2	88.0	78.9

Table 1: Progressive compression of XLM-R-base. Stages are grouped by degradation level. Highlighted rows indicate the baseline (gray) and optimal compression point (green, 85% reduction with 2.5% drop). Maximum compression rows (red) show higher degradation rates (7.6% drop). All F1 scores are averaged over 3 independent runs with different random seeds mBERT in Appendix J.

embedding size to 564, 456, and 312, truncating it to the first k dimensions. Training is performed under the supervision of the student from the previous stage. We find that using the original teacher leads to worse results, possibly due to the bigger knowledge gap (Wang et al., 2023). We also tested SVD-based dimensionality reduction but found truncation to be more effective (see Appendix C).

Vocabulary trimming: Restricting the vocabulary to the top 40K most frequent tokens for each target language introduces no measurable performance loss compared to the previous step, while further improving efficiency. Reducing below 40K works for some languages but does not generalize well across all cases (Appendix E), consistent with Ushio et al. (2023).

4.4 Downstream Performance

Our results show that model compression through knowledge distillation, structured pruning, and vocabulary reduction leads to modest performance drops (Tables 1 and 6). Below, we report results for XLM-R; results for mBERT follow similar patterns and are presented in Appendix J.

Language-specific resilience: The extent of degradation varies by language and correlates with teacher model quality. At maximum compression (92% parameter reduction), Slovak (1032MB fine-tuning (FT) data) experiences only a 2.9% performance drop, Swahili (332MB) shows a 5.2% drop, while Maltese (188MB) degrades by 19.2%. This pattern demonstrates that stronger teacher models–trained on larger datasets–enable more robust compression outcomes.

Task-specific patterns: Different tasks exhibit varying compression sensitivities. POS tagging shows the highest resilience across all languages, with performance drops of only 4-13% at 92% compression. Conversely, NER demonstrates steeper degradation, particularly for Maltese (69.2 $\rightarrow$ 37.1 F1). This severe drop is likely compounded by the extremely small Maltese NER training set (100 examples vs. 20,000 for Slovak), indicating that sequence labeling tasks are especially vulnerable to compression in low-resource settings. In contrast, sentence-level classification tasks such as SA and TC remain relatively stable under heavy compression, with performance decreases below 10% even at 85–90% size reduction.

Optimal compression trade-offs: The 85% compression level (hidden size 564 with 40k vocabulary) offers the best balance for most scenarios, with only a 2.5% average performance drop (84.3 vs 86.5 avg F1). For high-resource languages like Slovak, even 87% compression incurs only a 3.8% drop. Notably, vocabulary trimming often yields slight improvements (e.g., Maltese TC: 84.11 vs 83.43 F1), suggesting it reduces vocabulary noise while compensating for hidden size reduction.

Staged compression effects: Layer reduction (15%) and intermediate size pruning (20%) induce minimal degradation (<2% drop), with the primary performance impact occurring during hidden size reduction. Performance degrades gradually up to 85% compression, but deteriorates more rapidly beyond this threshold (4-6% drop per additional

stage).

Adapter capacity: We experiment with varying the reduction factor r to adjust task adapter capacity (Appendix I, Figure 10). While $r = 16$ suffices for larger models, smaller models (hidden sizes 564, 456, 312) benefit from lower r values ($r = 2$), yielding modest performance gains. Results in Tables 1 and 6 use $r = 2$ for these compressed models.

5 Related Work

In knowledge distillation, a smaller student model is trained to replicate the behavior of a larger teacher model (Hinton et al., 2015), often combining MLM loss with teacher supervision (Sun et al., 2019; Sanh et al., 2020). DistilBERT (Sanh et al., 2020) reduces model size by selecting every other layer from BERT (Devlin et al., 2019) and distills on large corpora using dynamic masking. Patient distillation further improves results by matching intermediate representations (Sun et al., 2019).

Recent work has explored distilling multilingual models into compact monolingual models. Singh and Lefever (2022) train student models for languages such as Swahili and Slovenian using a composite loss (distillation, cosine, MLM), and show that distilled models often outperform mBERT while using a reduced vocabulary (Abdaoui et al., 2020). Ansell et al. (2023) introduce a two-phase bilingual distillation pipeline, combining general-purpose and task-specific guidance with sparse fine-tuning, outperforming multilingual baselines.

Other studies emphasize the role of initialization. Wibowo et al. (2024) show that copying teacher weights is more effective than random initialization in the context of multilingual distillation, and that MSE outperforms KL divergence for distillation. Cruz (2025) similarly distill mBERT for Tagalog and highlight the nuanced impact of embedding initialization.

6 Conclusion

We present an effective compression pipeline for multilingual encoder models designed for low-resource languages. By integrating staged knowledge distillation, structured pruning, hidden size truncation, and vocabulary reduction, we compress models by up to 92% while maintaining competitive performance, typically within 2–10% of the original for moderate compression and 8–13% at maximum compression, on four downstream tasks.

Limitations

Our evaluation is limited to three low-resource languages and four downstream tasks, which may affect generalizability to other languages and task types. The compression pipeline requires target-language data for teacher adaptation, making it less suitable for truly low-resource languages with minimal corpora. We focus exclusively on encoder-only models (mBERT and XLM-R), and our structured pruning only targets feed-forward layers, leaving attention head pruning unexplored due to performance degradation.

Acknowledgments

This research was supported by DisAI - Improving scientific excellence and creativity in combating disinformation with artificial intelligence and language technologies, a Horizon Europe-funded project under GA No. 101079164, by the German Ministry of Education and Research (BMBF) as part of the project TRAILS (01IW24005), and by lorAI - Low Resource Artificial Intelligence, a project funded by the European Union under GA No.101136646.

References

Amine Abdaoui, Camille Pradel, and Grégoire Sigel. 2020. Load what you need: Smaller versions of mutililingual BERT. In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 119–123, Online. Association for Computational Linguistics.

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

Alan Ansell, Edoardo Maria Ponti, Anna Korhonen, and Ivan Vulić. 2023. Distilling efficient language-specific models for cross-lingual transfer.

Akari Asai, Sneha Kudugunta, Xinyan Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi. 2024. BUFFET: Benchmarking large language models for few-shot cross-lingual transfer. In *Proceedings of*

the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 1771–1800, Mexico City, Mexico. Association for Computational Linguistics.

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

Keith Cortis and Brian Davis. 2019. A social opinion gold standard for the Malta government budget 2018. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 364–369, Hong Kong, China. Association for Computational Linguistics.

Jan Christian Blaise Cruz. 2025. Extracting general-use transformers for low-resource languages via knowledge distillation. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 219–224, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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. Association for Computational Linguistics.

Alexiei Dingli and Nicole Sant. 2016. Sentiment analysis on maltese using machine learning. In *Proceedings of The Tenth International Conference on Advances in Semantic Processing (SEMAPRO 2016)*, pages 21–25.

Cheikh M. Bamba Dione, David Ifeoluwa Adelani, Peter Nabende, Jesujoba Alabi, Thapelo Sindane, Happy Buzaaba, Shamsuddeen Hassan Muhammad, Chris Chinenye Emezue, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jonathan Mukiibi, Blessing Sibanda, Bonaventure F. P. Dossou, Andiswa Bukula, Rooweither Mabuya, Allahsera Auguste Tapo, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Fatoumata Ouoba Kabore, Amelia Taylor, Godson Kalipe, Tebogo Macucwa, Vukosi Marivate, Tajuddeen Gwadabe, Mboning Tchiaze Elvis, Ikechukwu Onyenwe, Gratien Atindogbe, Tolulope Adelani, Idris Akinade, Olanrewaju Samuel, Marien Nahimana, Théogène Musabeyezu, Emile Niyomutabazi, Ester Chimhenga, Kudzai Gotosa, Patrick Mizha, Apelete Agbolo, Seydou Traore, Chinedu Uchechukwu, Aliyu Yusuf, Muhammad Abdullahi, and Dietrich Klakow. 2023. MasakhaPOS: Part-of-speech tagging for typologically diverse African languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10883–10900, Toronto, Canada. Association for Computational Linguistics.

Daniil Gurgurov, Ivan Vykopal, Josef van Genabith, and Simon Ostermann. 2025. Small models, big impact: Efficient corpus and graph-based adaptation of small multilingual language models for low-resource languages.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network.

Lu Hou, Zhiqi Huang, Lifeng Shang, Xin Jiang, Xiao Chen, and Qun Liu. 2020. Dynabert: Dynamic bert with adaptive width and depth. *Advances in Neural Information Processing Systems*, 33:9782–9793.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multitask benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Amir Hossein Kargaran, François Yvon, and Hinrich Schütze. 2025. Glotcc: An open broad-coverage commoncrawl corpus and pipeline for minority languages.

Taehyeon Kim, Jaehoon Oh, NakYil Kim, Sangwook Cho, and Se-Young Yun. 2021. Comparing kullback-leibler divergence and mean squared error loss in knowledge distillation.

Young Jin Kim and Hany Hassan. 2020. FastFormers: Highly efficient transformer models for natural language understanding. In *Proceedings of SustaiNLP:*

Workshop on Simple and Efficient Natural Language Processing, pages 149–158, Online. Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Felermino D'ario M'ario Ant'onio Ali, Davis Davis, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Gwadabe, Samuel Rutunda, Tadesse Belay, Wendimu Baye Messelle, Hailu Beshada Balcha, Sisay Adugna Chala, Hagos Tesfahun Gebremichael, Bernard Opoku, and Steven Arthur. 2023a. Afrisenti: A twitter sentiment analysis benchmark for african languages.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Seid Muhie Yimam, David Ifeoluwa Adelani, Ibrahim Sa'id Ahmad, Nedjma Ousidhoum, Abinew Ayele, Saif M Mohammad, and Meriem Beloucif. 2023b. Semeval-2023 task 12: Sentiment analysis for african languages (afrisenti-semeval). *arXiv preprint arXiv:2304.06845*.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Made Nindyatama Nityasya, Haryo Akbarianto Wibowo, Rendi Chevi, Radityo Eko Prasojo, and Alham Fikri Aji. 2022. Which student is best? a comprehensive knowledge distillation exam for task-specific bert models.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Samuel Pecar, Marian Simko, and Maria Bielikova. 2019. Improving sentiment classification in Slovak language. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 114–119, Florence, Italy. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Pranaydeep Singh and Els Lefever. 2022. When the student becomes the master: Learning better and smaller monolingual models from mBERT. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4434–4441, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz. 2006. Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In *Australasian joint conference on artificial intelligence*, pages 1015–1021. Springer.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2020. Energy and policy considerations for modern deep learning research. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(09):13693–13696.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient knowledge distillation for bert model compression.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

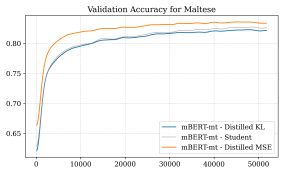
Asahi Ushio, Yi Zhou, and Jose Camacho-Collados. 2023. Efficient multilingual language model compression through vocabulary trimming. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14725–14739, Singapore. Association for Computational Linguistics.

Maorong Wang, Hao Yu, Ling Xiao, and Toshihiko Yamasaki. 2023. Bridging the capacity gap for online knowledge distillation. In *2023 IEEE 6th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 1–4.

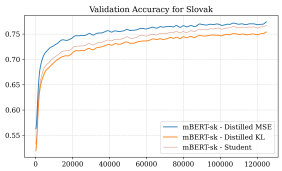
Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Haryo Akbarianto Wibowo, Thamar Solorio, and Alham Fikri Aji. 2024. The privileged students: On the value of initialization in multilingual knowledge distillation.

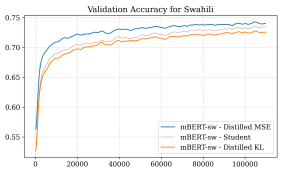
A KL Divergence vs MSE for Knowledge Distillation



(a) Maltese

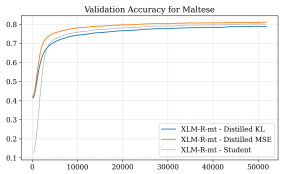


(b) Slovak

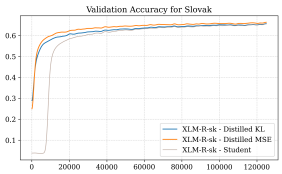


(c) Swahili

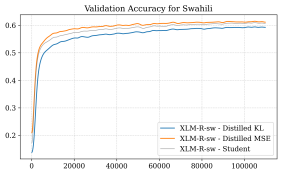
Figure 3: MSE vs. KD validation accuracy for mBERT with the models initialized using the last k layers.



(a) Maltese



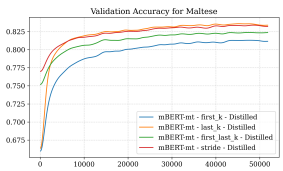
(b) Slovak



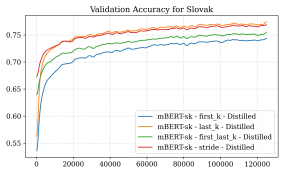
(c) Swahili

Figure 4: MSE vs. KD validation accuracy for XLM-R with the models initialized using the last k layers.

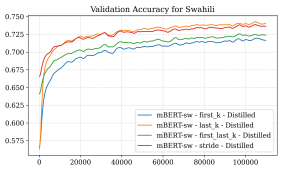
B Initialization Strategies for Knowledge Distillation



(a) Maltese

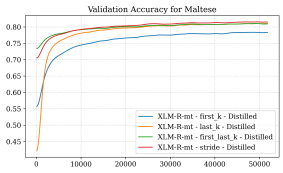


(b) Slovak

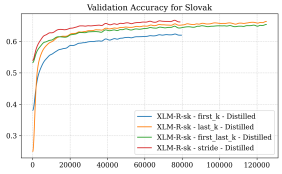


(c) Swahili

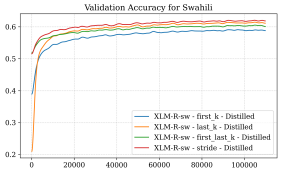
Figure 5: Validation accuracy for various initialization strategies for mBERT.



(a) Maltese



(b) Slovak



(c) Swahili

Figure 6: Validation accuracy for various initialization strategies for XLM-R.

C SVD vs. Truncation for Hidden Size Reduction

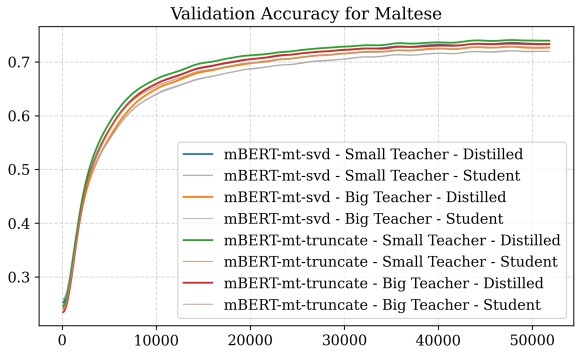
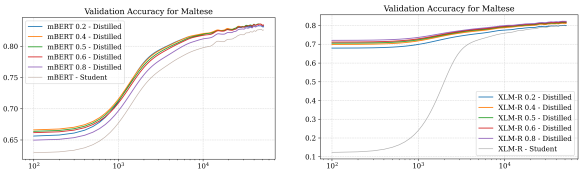


Figure 7: Validation accuracy comparing SVD vs. first-k truncation for hidden size reduction to 312. "Small teacher" refers to the layer-compressed (6-layer) model; "Big teacher" is the original 12-layer language-adapted model. Truncation consistently outperforms SVD regardless of teacher size.

D Alpha Parameter in Knowledge Distillation

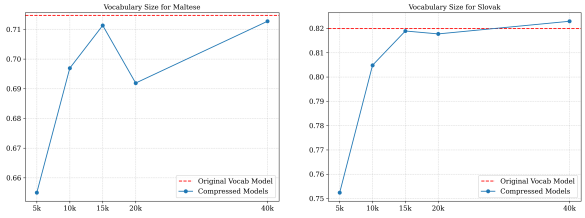


(a) mBERT (b) XLM-R

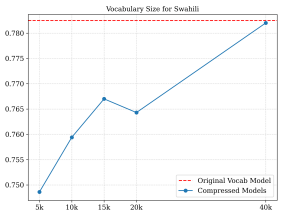
Figure 8: Validation accuracy curves showing the impact of the alpha parameter on knowledge distillation performance for mBERT and XLM-R on Maltese with the last k and stride initialization strategies for the two models respectively.

We find that the α parameter does not have a significant impact on mBERT during pre-training, with $\alpha = 0.5$ yielding consistently good results. For XLM-R, higher values of α (i.e., 0.6 and 0.8), which reduce the strength of the distillation effect, show slightly improved validation accuracy trends compared to lower values. In our experiments, we adopt the default setting of $\alpha = 0.5$, leaving a more comprehensive exploration of optimal values across different languages, dataset sizes, and model architectures to future work.

E Vocabulary Reduction Analysis



(a) Maltese (b) Slovak



(c) Swahili

Figure 9: Impact of vocabulary reduction on TC performance for mBERT models reduced to a hidden size of 312.

F Knowledge Distillation Data Sizes

Language	KD Data Size (MB)	FT Data Size (MB)
Maltese (mt)	238	188
Slovak (sk)	535	1032
Swahili (sw)	402	332

Table 2: Dataset sizes for knowledge distillation (KD) and monolingual fine-tuning (FT) for each language. The language-adapted models are sourced from Gurgurov et al. (2025), and the FT data sizes are as reported by them.

G Downstream Task Data Sizes

Language	Train	Validation	Test
Text Classification (TC)			
Maltese (mt)	701	99	204
Slovak (sk)	701	99	204
Swahili (sw)	701	99	204
Sentiment Analysis (SA)			
Maltese (mt)	595	85	171
Slovak (sk)	3560	522	1042
Swahili (sw)	738	185	304
Named Entity Recognition (NER)			
Maltese (mt)	100	100	100
Slovak (sk)	20000	10000	10000
Swahili (sw)	1000	1000	1000
Part of Speech Tagging (POS)			
Maltese (mt)	1123	433	518
Slovak (sk)	8483	1060	1061
Swahili (sw)	675	134	539

Table 3: Fine-tuning data sizes for each task (Text Classification, Sentiment Analysis, Named Entity Recognition, Part of Speech Tagging) showing train, validation, and test splits across Maltese, Slovak, and Swahili.

H Downstream Task Hyperparameters

Hyperparameter	TC	SA	NER	POS
Learning rate	1e-4	1e-4	3e-4	3e-4
Batch size	16	16	64	64
Epochs	20	20	100	100
Maximum length	256	256	512	512

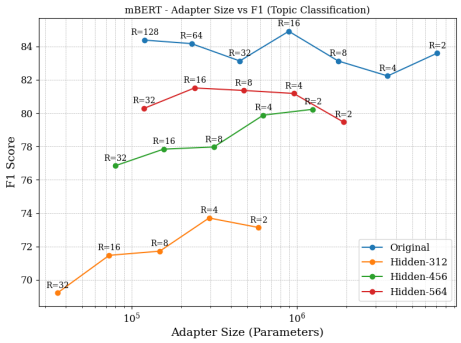
Table 4: Hyperparameters for task adapter fine-tuning across Text Classification (TC), Sentiment Analysis (SA), and Named Entity Recognition (NER) tasks.

I Adapter Trainable Parameter Counts

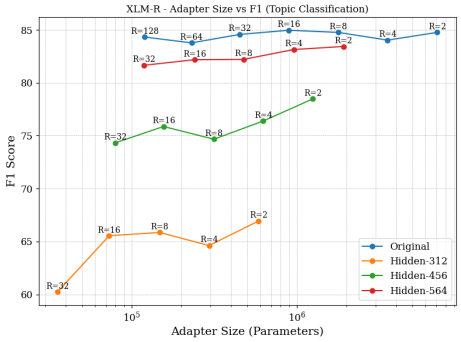
To examine whether the constrained task adapter capacity, as shown in Table 5, impacts downstream performance in compressed models, we vary the reduction factor r, thereby increasing adapter size (see Figure 10). We train task adapters on top of both full adapted models and hidden-size reduced models (564, 456, and 312). For the smallest models (456 and 312), we observe that increasing adapter capacity (r=2) leads to improved performance. However, this increase is unnecessary for larger mBERT variants (full and 564), while still beneficial for all small XLM-R models. These results suggest that for smaller models, increasing adapter capacity can yield modest performance gains. Tables 1 and 6 report results using the default reduction rate of 16.

Model Configuration	Task Adapter Size	
	mBERT	XLM-R
Base	894,528	894,528
Base-[mt, sk, sw]	894,528	894,528
KD layer red. ×2	447,264	447,264
inter. layer red. → 2048	447,264	447,264
* KD hid. size red. → 564	240,474	240,474
vocab. red. → 40k	240,474	240,474
* KD hid. size red. → 456	156,120	156,120
vocab. red. → 40k	156,120	156,120
* KD hid. size red. → 312	73,122	73,122
vocab. red. → 40k	73,122	73,122

Table 5: Task adapter parameter sizes across different model compression configurations for mBERT and XLM-R **with the default reduction factor of 16**. When the hidden size is reduced, adapter input/output dimensions decrease proportionally. When the layer count is reduced, fewer adapters are added to the model. All other parameters use the default settings for the Sequential Bottleneck adapter as implemented in AdapterHub.



(a) mBERT



(b) XLM-R

Figure 10: Performance of models on TC for Maltese with varying adapter capacity for mBERT and XLM-R.

J Downstream Results for mBERT

Compression Stage	Params	Size	Task Performance (F1)												Avg
			Maltese				Slovak				Swahili				
			TC	SA	NER	POS	TC	SA	NER	POS	TC	SA	NER	POS	
Baselines															
Multilingual	179M	179M	68.7	65.8	60.0	89.0	85.3	92.0	91.4	97.0	69.6	64.6	83.8	87.6	79.6
Language-adapted	179M	179M	84.9	73.6	65.0	94.0	86.3	91.9	90.4	96.9	86.7	81.3	82.5	88.7	**85.2**
Compression Pipeline (minimal degradation)															
Layer reduction	135M (-25%)	135M	80.1	73.9	59.0	93.2	85.4	90.4	87.4	96.9	82.8	77.3	80.7	88.4	83.0
+ FFN pruning	126M (-30%)	126M	79.0	74.7	58.1	92.7	85.3	90.2	88.5	96.7	83.2	75.9	79.8	88.5	82.7
+ Hidden 564	90M (-50%)	90M	79.5	70.2	61.1	92.6	83.4	90.5	88.1	96.3	83.5	76.1	79.7	88.4	82.5
+ Vocabulary	45M (-75%)	45M	80.2	70.8	61.1	92.5	83.5	90.7	87.7	96.3	84.3	76.0	80.3	88.6	**82.7**
Further compression (moderate degradation)															
+ Hidden 456	71M (-60%)	71M	80.2	70.1	57.2	92.1	83.9	90.4	87.5	95.9	85.1	78.6	80.3	88.3	82.5
+ Vocabulary	35M (-80%)	35M	81.0	69.7	55.9	92.0	84.2	90.4	87.4	96.0	83.0	78.5	79.8	88.4	82.2
Maximum compression (higher degradation)															
+ Hidden 312	48M (-73%)	48M	73.1	72.0	39.5	90.3	80.9	90.4	86.5	95.5	81.8	76.5	79.6	87.7	79.5
+ Vocabulary	23M (-87%)	23M	73.0	72.1	40.4	90.2	81.9	90.1	86.2	95.3	81.7	76.0	77.1	87.7	79.3

Table 6: Progressive compression of mBERT. Stages are grouped by degradation level. Highlighted rows indicate the baseline (gray) and optimal compression point (green, 75% reduction with 2.5% drop). Maximum compression rows (red) show significant degradation (5.9% drop). TC=Topic Classification, SA=Sentiment Analysis, NER=Named Entity Recognition, POS=Part-of-Speech Tagging. F1 scores averaged over 3 runs.