Adapting Pre-trained Language Models to African Languages via Multilingual Adaptive Fine-Tuning

Jesujoba O. Alabi,* **David Ifeoluwa Adelani,*** **Marius Mosbach, and Dietrich Klakow**
Spoken Language Systems (LSV), Saarland University, Saarland Informatics Campus, Germany
{jalabi,didelani,mmosbach,dklakow}@lsv.uni-saarland.de

Abstract

Multilingual pre-trained language models (PLMs) have demonstrated impressive performance on several downstream tasks for both high-resourced and low-resourced languages. However, there is still a large performance drop for languages unseen during pre-training, especially African languages. One of the most effective approaches to adapt to a new language is *language adaptive fine-tuning* (LAFT) — fine-tuning a multilingual PLM on monolingual texts of a language using the pre-training objective. However, adapting to a target language individually takes a large disk space and limits the cross-lingual transfer abilities of the resulting models because they have been specialized for a single language. In this paper, we perform *multilingual adaptive fine-tuning* on 17 most-resourced African languages and three other high-resource languages widely spoken on the African continent to encourage cross-lingual transfer learning. To further specialize the multilingual PLM, we removed vocabulary tokens from the embedding layer that corresponds to non-African writing scripts before MAFT, thus reducing the model size by around 50%. Our evaluation on two multilingual PLMs (AfriBERTa and XLM-R) and three NLP tasks (NER, news topic classification, and sentiment classification) shows that our approach is competitive to applying LAFT on individual languages while requiring significantly less disk space. Additionally, we show that our adapted PLM also improves the zero-shot cross-lingual transfer abilities of parameter efficient fine-tuning methods.

1 Introduction

Recent advances in the development of multilingual pre-trained language models (PLMs) like mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), and RemBERT (Chung et al., 2021) have led to significant performance gains on a wide range of cross-lingual

transfer tasks. Due to the *curse of multilinguality* (Conneau et al., 2020) — a trade-off between language coverage and model capacity — and non-availability of pre-training corpora for many low-resource languages, multilingual PLMs are often trained on about 100 languages. Despite the limitations of language coverage, multilingual PLMs have been shown to transfer to several low-resource languages unseen during pre-training. Although, there is still a large performance gap compared to languages seen during pre-training.

One of the most effective approaches to adapt to a new language is *language adaptive fine-tuning* (LAFT) — fine-tuning a multilingual PLM on monolingual texts in the target language using the same pre-training objective. This has been shown to lead to big gains on many cross-lingual transfer tasks (Pfeiffer et al., 2020), and low-resource languages (Muller et al., 2021; Chau & Smith, 2021), including African languages (Alabi et al., 2020; Adelani et al., 2021). Nevertheless, adapting a model to each target language individually takes large disk space, and limits the cross-lingual transfer abilities of the resulting models because they have been specialized to individual languages (Beukman, 2021).

An orthogonal approach to improve the coverage of low-resource languages is to include them in the pre-training data. An example for this approach is AfriBERTa (Ogueji et al., 2021), which was trained from scratch on 11 African languages. A downside of this approach is that it is resource intensive in terms of data and compute.

Another alternative approach is parameter efficient fine-tuning like Adapters (Pfeiffer et al., 2020) and sparse fine-tuning (Ansell et al., 2021), where the model is adapted to new languages by using a sparse network trained on a small monolingual corpus. Similar to LAFT, it requires adaptation for every new target language. Although it takes little disk space, all target language-specific parameters need to be stored.

In this paper, we propose *multilingual adaptive fine-tuning (MAFT)*, a language adaptation to multiple languages at once. We perform language adaptation on the 17 most-resourced African languages (Afrikaans, Amharic, Hausa, Igbo, Malagasy, Chichewa, Oromo, Naija, Kinyarwanda, Kirundi, Shona, Somali, Sesotho, Swahili, isiXhosa, Yorùbá, isiZulu) and three other high-resource language widely spoken on the continent

---

* Equal contribution.

(English, French, and Arabic) *simultaneously* to provide a single model for cross-lingual transfer learning for African languages. To further specialize the multilingual PLM, we follow the approach of Abdaoui et al. (2020) and remove vocabulary tokens from the embedding layer that correspond to non-Latin and non-Ge'ez (used by Amharic) scripts before MAFT, thus effectively reducing the model size by 50%.

Our evaluation on two multilingual PLMs (AfriBERTa and XLM-R) and three NLP tasks (NER, news topic classification and sentiment classification) shows that our approach is competitive to performing LAFT on the individual languages, with the benefit of having a single model instead of a separate model for each of the target languages. Also, we show that our adapted PLM improves the zero-shot cross-lingual transfer abilities of parameter efficient fine-tuning methods like Adapters (Pfeiffer et al., 2020) and sparse fine-tuning (Ansell et al., 2021).

As an additional contribution, and in order to cover more diverse African languages in our evaluation, we create a new evaluation corpus, ANTC – **A**frican **N**ews **T**opic **C**lassification – for Lingala, Somali, Naija, Malagasy, and isiZulu from pre-defined news categories of VOA, BBC, Global Voices, and Isolezwe newspapers. To further the research on NLP for African languages, we make our code and data publicly available.[1] Additionally, our models are available via HuggingFace.[2]

2 Related Work

Multilingual PLMs for African languages The success of multilingual PLMs such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) for cross-lingual transfer in many natural language understanding tasks has encouraged the continuous development of multilingual models (Luo et al., 2021; Chi et al., 2021; Ouyang et al., 2021; Chung et al., 2021; He et al., 2021). Most of these models cover 50 to 110 languages and only few African languages are represented due to lack of large monolingual corpora on the web. To address this under-representation, regional multilingual PLMs have been trained from scratch such as AfriBERTa (Ogueji et al., 2021) or adapted from existing multilingual PLM through LAFT (Alabi et al., 2020; Pfeiffer et al., 2020; Muller et al., 2021; Adelani et al., 2021). AfriBERTa is a relatively small multilingual PLM (126M parameters) trained using the RoBERTa architecture and pre-training objective on 11 African languages. However, it lacks coverage of languages from the southern region of the African continent, specifically the southern-Bantu languages. In our work, we extend to those languages since only a few of them have large (>100MB size) monolingual corpus. We also do not specialize to a single language

but apply MAFT which allows multilingual adaptation and preserves downstream performance on both high-resource and low-resource languages.

Adaptation of multilingual PLMs It is not unusual for a new multilingual PLM to be initialized from an existing model. For example, Chi et al. (2021) trained InfoXLM by initializing the weights from XLM-R before training the model on a joint monolingual and translation corpus. Although they make use of a new training objective during adaptation. Similarly, Tang et al. (2020) extended the languages covered by mBART (Liu et al., 2020b) from 25 to 50 by first modifying the vocabulary and initializing the model weights of the original mBART before fine-tuning it on a combination of monolingual texts from the original 25 languages in addition to 25 new languages. Despite increasing the number of languages covered by their model, they did not observe a significant performance drop on downstream tasks. We take inspiration from these works for applying MAFT on African languages, but we do not modify the training objective during adaptation nor increase the vocabulary.

Compressing PLMs One of the most effective methods for creating smaller PLMs is distillation where a small student model is trained to reproduce the behaviour of a larger teacher model. This has been applied to many English PLMs (Sanh et al., 2019; Jiao et al., 2020; Sun et al., 2020; Liu et al., 2020a) and a few multilingual PLMs (Wang et al., 2020, 2021). However, it often leads to a drop in performance compared to the teacher PLM. An alternative approach that does not lead to a drop in performance has been proposed by Abdaoui et al. (2020) for multilingual PLM. They removed unused vocabulary tokens from the embedding layer. This simple method significantly reduces the number of embedding parameters thus reducing the overall model size since the embedding layer contributes the most to the total number of model parameters. In our paper, we combine MAFT with the method proposed by Abdaoui et al. (2020) to reduce the overall size of the resulting multilingual PLM for African languages. This is crucial especially because people from under-represented communities in Africa may not have access to powerful GPUs in order to fine-tune large PLMs. Also, Google Colab[3] (free-version), which is widely used by individuals from under-represented communities without access to other compute resources, cannot run large models like e.g. XLM-R. Hence, it is important to provide smaller models that still achieve competitive downstream performance to these communities.

Evaluation datasets for African languages One of the challenges of developing (multilingual) PLMs for African languages is the lack of evaluation corpora. There have been many efforts by communities like

---

[1] https://github.com/uds-lsv/afro-maft
[2] https://huggingface.co/Davlan

[3] https://colab.research.google.com/

Domain	Number of sentences			Classes	Number of classes
	Train	Dev	Test		
Newly created datasets					
Lingala (lin)	1,536	220	440	Rdc, Politiki/Politique, Bokengi/Securite, Justice, Bokolongono/Santé/Medecine	5
Naija (pcm)	1,165	167	333	Entertainment, Africa, Sport, Nigeria, World	5
Malagasy (mlg)	4544	650	1299	Politika (Politics), Kolontsaina (Culture), Zon'olombelona (Human Rights), Siansa_sy_Teknolojia (Science and Technology) ,Tontolo_iainana (Environment)	5
Somali (som)	10,072	1,440	2879	Soomaaliya (Somalia), Wararka (News), Caalamka (World), Maraykanka (United States), Afrika (Africa)	6
isiZulu (zul)	2,961	424	847	Ezemidlalo (Sports), Ezokungcebeleka (Recreation), Imibono (Ideas), Ezezimoto (Automotive), Intandokazi (Favorites)	5
Existing datasets					
Amharic (amh)	36,029	5,147	10,294	Local News, Sport, Politics, International News, Business, Entertainment	6
English (eng)	114,000	6,000	7,600	World, Sports, Business, Sci/Tech	4
Hausa (hau)	2,045	290	582	Africa, World, Health, Nigeria, Politics	5
Kinyarwanda (kin)	16,163	851	4,254	Politics, Sport, Economy, Health, Entertainment, History, Technology, Tourism, Culture, Fashion, Religion, Environment, Education, Relationship	14
Kiswahili (swa)	21,096	1,111	7,338	Uchumi (Economic), Kitaifa (National), Michezo (Sports), Kimataifa (International), Burudani (Recreation), Afya (Health)	6
Yorùbá (yor)	1,340	189	379	Nigeria, Africa, World, Entertainment, Health, Sport, Politics	7

Table 1: Number of sentences in training, development and test splits. We provide automatic translation of some of the African language words to English (in Parenthesis) using Google Translate.

Masakhane to address this issue (∀ et al., 2020; Adelani et al., 2021). We only find two major evaluation benchmark datasets that cover a wide range of African languages: one for named entity recognition (NER) (Adelani et al., 2021) and one for sentiment classification (Muhammad et al., 2022). In addition, there are also several news topic classification datasets (Hedderich et al., 2020; Niyongabo et al., 2020; Azime & Mohammed, 2021) but they are only available for a few African languages. Our work contributes novel news topic classification datasets (i.e. ANTC) for additional five African languages: Lingala, Naija, Somali, isiZulu, and Malagasy.

3 Data

3.1 Adaptation corpora

We perform MAFT on 17 African languages Afrikaans, Amharic, Hausa, Igbo, Malagasy, Chichewa, Oromo, Naija, Kinyarwanda, Kirundi, Shona, Somali, Sesotho, Swahili, isiXhosa, Yorùbá, isiZulu) covering the major African language families and 3 high resource languages (Arabic, French, and English) widely spoken in Africa. We selected the African languages based on the availability of a (relatively) large amount of monolingual texts. We obtain the monolingual texts from three major sources: the mT5 pre-training corpus which is based on Common Crawl Corpus[4] (Xue et al., 2021), the British Broadcasting Corporation (BBC)

News, Voice of America News[5] (Palen-Michel et al., 2022), and some other news websites based in Africa. Table 9 in the Appendix provides a summary of the monolingual data, including their sizes and sources. We pre-processed the data by removing lines that consist of numbers or punctuation only, and lines with less than six tokens.

3.2 Evaluation tasks

We run our experiments on two sentence level classification tasks: news topic classification and sentiment classification, and one token level classification task: NER. We evaluate our models on English as well as diverse African languages with different linguistic characteristics.

3.2.1 Existing datasets

NER For the NER task we evaluate on the MasakhaNER dataset (Adelani et al., 2021), a manually annotated dataset covering 10 African languages (Amharic, Hausa, Igbo, Kinyarwanda, Luganda, Luo, Naija, Kiswahili, Wolof, and Yorùbá) with texts from the news domain. For English, we use data from the CoNLL 2003 NER task (Tjong Kim Sang & De Meulder, 2003) also containing texts from the news domain. For isiXhosa, we use the data from Eiselen (2016). Lastly, to evaluate on Arabic we make use of the ANERCorp dataset (Benajiba et al., 2007; Obeid et al., 2020).

[4] https://commoncrawl.org/
[5] https://www.voanews.com

News topic classification We use existing news topic datasets for Amharic (Azime & Mohammed, 2021), English – AG News corpus – (Zhang et al., 2015), Kinyarwanda – KINNEWS – (Niyongabo et al., 2020), Kiswahili – new classification dataset– (David, 2020), and both Yorùbá and Hausa (Hedderich et al., 2020). For dataset without a development set, we randomly sample 5% of their training instances and use them as a development set.

Sentiment classification We use the NaijaSenti multilingual Twitter sentiment analysis corpus (Muhammad et al., 2022). This is a large code-mixed and monolingual sentiment analysis dataset, manually annotated for 4 Nigerian languages: Hausa, Igbo, Yorùbá and Pidgin. Additionally, we evaluate on the Amharic, and English Twitter sentiment datasets by Yimam et al. (2020) and Rosenthal et al. (2017), respectively. For all datasets above, we only make use of tweets with positive, negative and neutral sentiments.

3.2.2 Newly created dataset: ANTC corpus

We created a novel dataset, ANTC — **A**frican **N**ews **T**opic **C**lassification for five African languages. We obtained data from three different news sources: VOA, BBC[6], Global Voices[7], and isolezwe[8]. From the VOA data we created datasets for Lingala and Somali. We obtained the topics from data released by Palen-Michel et al. (2022) and used the provided URLs to get the news category from the websites. For Naija, Malagasy and isiZulu, we scrapped news topic from the respective news website (BBC Pidgin, Global Voices, and isolezwe respectively) directly base on their category. We noticed that some news topics are not mutually exclusive to their categories, therefore, we filtered such topics with multiple labels. Also, we ensured that each category has at least 200 samples. The categories include but are not limited to: Africa, Entertainment, Health, and Politics. The pre-processed datasets were divided into training, development, and test sets using stratified sampling with a ratio of 70:10:20. Table 1 provides details about the dataset size and news topic information.

4 Pre-trained Language Models

For our experiments, we make use of different multilingual PLMs that have been trained using a masked language model objective on large collections of monolingual texts from several languages. Table 2 shows the number of parameters as well as the African languages covered by each of the models we consider.

1. XLM-R (Conneau et al., 2020) has been pre-trained on 100 languages including eight African languages. We make use of both XLM-R-base and XLM-R-large for MAFT with 270M and 550M

---

[6]https://www.bbc.com/pidgin
[7]https://mg.globalvoices.org/
[8]https://www.isolezwe.co.za

PLM	# Lang.	African languages covered
XLM-R-base (270M)	100	afr, **amh**, **hau**, **mlg**, orm, **som**, **swa**, **xho**
AfriBERTa-large (126M)	11	**amh**, **hau**, **ibo**, **kin**, run, orm, **pcm**, som, **swa**, tir, **yor**
XLM-R-miniLM (117M)	100	afr, **amh**, **hau**, **mlg**, orm, **som**, **swa**, **xho**
XLM-R-large (550M)	100	afr, **amh**, **hau**, **mlg**, orm, **som**, **swa**, **xho**
AfroXLMR* (117M-270M)	20	afr, **amh**, **hau**, **ibo**, **kin**, run **mlg**, nya, orm, **pcm**, **sna**, **som**, sot, **swa**, **xho**, **yor**, **zul**

Table 2: Language coverage and size for pre-trained language models. Languages in **bold** have evaluation datasets for either NER, news topic classification or sentiment analysis.

parameter sizes respectively. Although, for our main experiments, we make use of XLM-R-base.

2. AfriBERTa (Ogueji et al., 2021) has been pre-trained only on African languages. Despite its smaller parameter size (126M), it has been shown to reach competitive performance to XLM-R-base on African language datasets (Adelani et al., 2021; Hedderich et al., 2020).

3. XLM-R-miniLM (Wang et al., 2020) is a distilled version of XLM-R-large with only 117M parameters.

Hyper-parameters for baseline models We fine-tune the baseline models for NER, news topic classification and sentiment classification for 50, 25, and 20 epochs respectively. We use a learning rate of 5e-5 for all the task, except for sentiment classification where we use 2e-5 for XLM-R-base and XLM-R-large. The maximum sequence length is 164 for NER, 500 for news topic classification, and 128 for sentiment classification. The adapted models also make use of similar hyper-parameters.

5 Multilingual Adaptive Fine-tuning

We introduce MAFT as an approach to adapt a multilingual PLM to a new set of languages. Adapting PLMs has been shown to be effective when adapting to a new domain (Gururangan et al., 2020) or language (Pfeiffer et al., 2020; Alabi et al., 2020; Muller et al., 2021; Adelani et al., 2021). While previous work on multilingual adaptation has mostly focused on autoregressive sequence-to-sequence models such as mBART (Tang et al., 2020), in this work, we adapt non-autoregressive masked PLMs on monolingual corpora covering 20 languages. Crucially, during adaptation we use the same objective that was also used during pre-training. The models resulting from MAFT can then be fine-tuned on supervised NLP downstream tasks. We name the model resulting after applying MAFT to XLM-R-base and XLM-R-miniLM as *AfroXLMR-base* and *AfroXLMR-mini*, respectively. For adaptation, we train

Model	Size	amh	ara	eng	hau	ibo	kin	lug	luo	pcm	swa	wol	xho	yor	avg
Finetune															
XLM-R-miniLM	117M	69.5	76.1	91.5	74.5	81.9	68.6	64.7	11.7	83.2	86.3	51.7	69.3	72.0	69.3
AfriBERTa	126M	73.8	51.3	89.0	90.2	87.4	73.8	78.9	70.2	85.7	88.0	61.8	67.2	81.3	76.8
XLM-R-base	270M	70.6	77.9	92.3	89.5	84.8	73.3	79.7	74.9	87.3	87.4	63.9	69.9	78.3	79.2
XLM-R-large	550M	76.2	79.7	93.1	90.5	84.1	73.8	81.6	73.6	89.0	89.4	67.9	72.4	78.9	80.8
MAFT + Finetune															
XLM-R-miniLM	117M	69.7	76.5	91.7	87.7	83.5	74.1	77.4	17.5	85.5	86.0	59.0	72.3	75.1	73.5
AfriBERTa	126M	72.5	40.9	90.1	89.7	87.6	75.2	80.1	69.6	86.5	87.6	62.3	71.8	77.0	76.2
XLM-R-base	270M	76.1	79.7	**92.8**	91.2	87.4	78.0	82.9	75.1	89.6	88.6	67.4	71.9	82.1	81.8
XLM-R-base-v70k	140M	70.1	76.4	91.0	91.4	86.6	77.5	83.2	**75.4**	89.0	88.7	65.9	72.4	81.3	80.7
XLM-R-base+LAFT	270M x 13	**78.0**	79.1	91.3	**91.5**	87.7	77.8	**84.7**	75.3	**90.0**	**89.5**	**68.3**	**73.2**	**83.7**	**82.3**

Table 3: **NER** model comparison, showing F1-score on the test sets after 50 epochs averaged over 5 runs. Results are for all 4 tags in the dataset: PER, ORG, LOC, DATE/MISC. For LAFT, we multiplied the size of XLM-R-base by the number of languages as LAFT results in a single model per language.

on a combination of the monolingual corpora used for AfriMT5 adaptation by Adelani et al. (2022). Details for each of the monolingual corpora and languages are provided in Appendix A.1.

Hyper-parameters for MAFT The PLMs were trained for 3 epochs with a learning rate of 5e-5 using huggingface transformers (Wolf et al., 2020). We use of a batch size of 32 for AfriBERTa and a batch size 10 for the other PLMs.

5.1 Vocabulary reduction

Multilingual PLMs come with various parameter sizes, the larger ones having more than hundred million parameters, which makes fine-tuning and deploying such models a challenge due to resource constraints. One of the major factors that contributes to the parameter size of these models is the embedding matrix whose size is a function of the vocabulary size of the model. While a large vocabulary size is essential for a multilingual PLM trained on hundreds of languages, some of the tokens in the vocabulary can be removed when they are irrelevant to the domain or language considered in the downstream task, thus reducing the vocabulary size of the model. Inspired by Abdaoui et al. (2020), we experiment with reducing the vocabulary size of the XLM-R-base model before adapting via MAFT. There are two possible vocabulary reductions in our setting: *(1) removal of tokens before MAFT or (2) removal of tokens after MAFT*. From our preliminary experiments, we find approach (1) to work better. We call the resulting model, *AfroXLMR-small*.

To remove non-African vocabulary sub-tokens from the pretrained XLM-base model, we concatenated the monolingual texts from 19 out of the 20 African languages together. Then, we apply sentencepiece to the Amharic monolingual texts, and concatenated texts separately using the original XLM-R-base tokenizer. The frequency of all the sub-tokens in the two separate monolingual corpora is computed, and we select the top-k most frequent tokens from the separate corpora. We used this separate sampling to ensure that a considerable number of Amharic sub-tokens are captured

in the new vocabulary, we justify the choice of this approach in Section 5.3. We assume that the top-k most frequent tokens should be representative of the vocabulary of the whole 20 languages. We chose $k = 52.000$ from the Amharic sub-tokens which covers 99.8% of the Amharic monolingual texts, and $k = 60.000$ which covers 99.6% of the other 19 languages, and merged them. In addition, we include the top 1000 tokens from the original XLM-R-base tokenizer in the new vocabulary to include frequent tokens that were not present in the new top-k tokens.[9] We note that our assumption above may not hold in the case of some very distant and low-resourced languages as well as when there are domain differences between the corpora used during adaptation and fine-tuning. We leave the investigation of alternative approaches for vocabulary compression for future work.

5.2 Results and discussion

5.2.1 Baseline results

For the baseline models (top rows in Tables 3, 4, and 5), we directly fine-tune on each of the downstream tasks in the target language: NER, news topic classification and sentiment analysis.

Performance on languages seen during pre-training For NER and sentiment analysis we find XLM-R-large to give the best overall performance. We attribute this to the fact that it has a larger model capacity compared to the other PLMs. Similarly, we find AfriBERTa and XLM-R-base to give better results on languages they have been pre-trained on (see Table 2), and in most cases AfriBERTa tends to perform better than XLM-R-base on languages they are both pre-trained on, for example amh, hau, and swa. However, when the languages are unseen by AfriBERTa (e.g. ara, eng, wol, lin, lug, luo, xho, zul), it performs much worse than XLM-R-base and in some cases even worse than the XLM-R-miniLM. This shows that it may be better

---

[9]This introduced just a few new tokens which are mostly English tokens to the new vocabulary. We end up with 70.609 distinct sub-tokens after combining all of them.

Model	Size	amh	eng	hau	kin	lin	mlg	pcm	som	swa	yor	zul	avg
Finetune													
XLM-R-miniLM	117M	70.4	94.1	77.6	64.2	41.2	42.9	67.6	74.2	86.7	68.8	56.9	67.7
AfriBERTa	126M	70.7	93.6	90.1	75.8	55.4	56.4	81.5	**79.9**	87.7	**82.6**	71.4	76.8
XLM-R-base	270M	71.1	94.1	85.9	73.3	56.8	54.2	77.3	78.8	87.1	71.1	70.0	74.6
XLM-R-large	550M	72.7	94.5	86.2	75.1	52.2	63.6	79.4	79.2	87.5	74.8	78.7	76.7
MAFT + Finetune													
XLM-R-miniLM	117M	69.5	94.1	86.7	72.0	51.7	55.3	78.1	77.7	87.2	74.0	60.3	73.3
AfriBERTa	126M	68.8	93.7	89.5	76.5	54.9	59.7	**82.2**	**79.9**	87.7	80.8	76.4	77.3
XLM-R-base	270M	71.9	**94.6**	88.3	**76.8**	**58.6**	64.7	78.9	79.1	87.8	80.2	**79.6**	78.2
XLM-R-base-v70k	140M	70.4	94.2	87.7	76.1	56.8	64.4	76.1	79.4	87.4	76.9	77.4	76.9
XLM-R-base+LAFT	270M x 11	**73.0**	94.3	**91.2**	76.0	56.9	**67.3**	77.4	79.4	**88.0**	79.2	79.5	**78.4**

Table 4: **News topic classification** model comparison, showing F1-score on the test sets after 25 epochs averaged over 5 runs. For LAFT, we multiplied the size of XLM-R-base by the number of languages.

Model	Size	amh	eng	hau	ibo	pcm	yor	avg
Finetune								
XLM-R-miniLM	117M	51.0	62.8	75.0	78.0	72.9	73.4	68.9
AfriBERTa-large	126M	51.7	61.8	81.0	**81.2**	75.0	80.2	71.8
XLM-R-base	270M	51.4	66.2	78.4	79.9	76.3	76.9	71.5
XLM-R-large	550M	52.4	67.5	79.3	80.8	77.6	78.1	72.6
MAFT+Finetune								
XLM-R-miniLM	117M	51.3	63.3	77.7	78.0	73.6	74.3	69.7
AfriBERTa	126M	53.6	63.2	81.0	80.6	74.7	80.4	72.3
XLM-R-base	270M	53.0	65.6	80.7	80.5	77.5	79.4	72.8
XLM-R-base-v70k	140M	52.2	65.3	80.6	81.0	77.4	78.6	72.5
XLM-R-base+LAFT	270M x 6	**55.0**	65.6	**81.5**	80.8	74.7	**80.9**	**73.1**

Table 5: **Sentiment classification** model comparison, showing F1 evaluation on test sets after 20 epochs, averaged over 5 runs. We obtained the results for the baseline model results of "hau", "ibo", "pcm", and "yor" from Muhammad et al. (2022). For LAFT, we multiplied the size of XLM-R-base by the number of languages as LAFT results in a single model per language.

to adapt to a new African language from a PLM that has seen numerous languages than one trained on a subset of African languages from scratch.

LAFT is a strong baseline The results of applying LAFT to the XLM-R-base model are shown in the last row of Tables 3, 4, and 5. We find that applying LAFT on each language individually provides a significant improvement in performance across all languages and tasks we evaluated on. Sometimes, the improvement is very large, for example, +7.4 F1 on Amharic NER and +9.5 F1 for Zulu news-topic classification. The only exception is for English since XLM-R has already seen large amounts of English text during pre-training. Additionally, LAFT models tend to give slightly worse result when adaptation is performed on a smaller corpus.[10]

5.2.2 Multilingual adaptive fine-tuning results

While LAFT provides an upper bound on downstream performance for most languages, our new approach is often competitive to LAFT. On average, the difference on NER, news topic and sentiment classification is −0.5, −0.2, and −0.3 F1, respectively. Crucially,

---
[10]We performed LAFT on eng using VOA news corpus with about 906.6MB, much smaller than the CC-100 eng corpus (300GB)

compared to LAFT, MAFT results in a single adapted model which can be applied to many languages while LAFT results in a new model for each language. Below, we discuss our results in more detail.

PLMs pre-trained on many languages benefit the most from MAFT We found all the PLMs to improve after we applied MAFT. The improvement is the largest for the XLM-R-miniLM, where the performance improved by +4.2 F1 for NER, and +5.6 F1 for news topic classification. Although, the improvement was lower for sentiment classification (+0.8). Applying MAFT on XLM-R-base gave the overall best result. On average, there is an improvement of +2.6, +3.6, and +1.5 F1 on NER, news topic and sentiment classification, respectively. The main advantage of MAFT is that it allows us to use the same model for many African languagesinstead of many models specialized to individual languages. This significantly reduces the required disk space to store the models, without sacrificing performance. Interestingly, there is no strong benefit of applying MAFT to AfriBERTa. In most cases the improvement is < 0.6 F1. We speculate that this is probably due to AfriBERTa's tokenizer having a limited coverage. We leave a more detailed investigation of this for future work.

Model	amh		ara		eng		yor	
	#UNK	F1	#UNK	F1	#UNK	F1	#UNK	F1
AfroXLMR-base	0	76.1	0	79.7	0	92.8	24	82.1
Afro-XLM-R70k (i)	3704	67.8	1403	76.3	44	90.6	5547	81.2
Afro-XLM-R70k (ii)	3395	70.1	669	76.4	54	91.0	6438	81.3

Table 6: Numbers of UNKs when the model tokenizers are applied on the NER test sets.

More efficient models using vocabulary reduction
Applying vocabulary reduction helps to reduce the model size by more than 50% before applying MAFT. We find a slight reduction in performance as we remove more vocabulary tokens. Average performance of XLM-R-base-v70k reduces by -1.6, -1.5 and -0.6 F1 for NER, news topic, and sentiment classification compared to the XLM-R-base+LAFT baseline. Despite the reduction in performance compared to XLM-R-base+LAFT, they are still better than XLM-R-miniLM, which has a similar model size, with or without MAFT. We also find that their performance is better than that of the PLMs that have not undergone any adaptation. We find the largest reduction in performance on languages that make use of non-Latin scripts i.e. amh and ara — they make use of the Ge'ez script and Arabic script respectively. We attribute this to the vocabulary reduction impacting the number of amh and ara subwords covered by our tokenizer.

In summary, we recommend XLM-R-base+MAFT (i.e. AfroXLMR-base) for all languages on which we evaluated, including high-resource languages like English, French and Arabic. If there are GPU resource constraints, we recommend using XLM-R-base-v70k+MAFT (i.e. AfroXLMR-small).

5.3 Ablation experiments on vocabulary reduction

Our results showed that applying vocabulary reduction reduced the model size, but we also observed a drop in performance for different languages across the downstream tasks, especially for Amharic, because it uses a non-Latin script. Hence, we compared different sampling strategies for selecting the top-k vocabulary sub-tokens. These include: (i) concatenating the monolingual texts, and selecting the top-70k sub-tokens (ii) the exact approach described in Section 5.1. The resulting tokenizers from the two approaches are used to tokenize the sentences in the NER test sets for Amharic, Arabic, English, and Yorùbá. Table 6 shows the number of UNKs in the respective test set after tokenization and the F1 scores obtained on the NER task for the languages. The table shows that the original AfroX-LMR tokenizer obtained the least number of UNKs for all languages, with the highest F1 scores. Note that Yorùbá has 24 UNKs, which is explained by the fact that Yorùbá was not seen during pre-training. Furthermore, using approach (i), gave 3704 UNKs for Amharic, but with approach (ii) there was a significant drop in the number of UNKs and an improvement in F1

score. We noticed a drop in the vocabulary coverage for the other languages as we increased the Amharic sub-tokens. Therefore, we concluded that there is no sweet spot in terms of the way to pick the vocabulary that covers all languages and we believe that this is an exciting area for future work.

5.4 Scaling MAFT to larger models

To demonstrate the applicability of MAFT to larger models, we applied MAFT to XLM-R-large using the same training setup as XLM-R-base. We refer to the new PLM as AfroXLMR-large. For comparison, we also trained individual LAFT models using the monolingual data[11] from Adelani et al. (2021). Table 7 shows the evaluation result on NER. Averaging over all 13 languages, AfroXLMR-large improved over XLM-R-large by $+2.8$ F1, which is very comparable to the improvement we obtained between AfroXLMR-base (81.8 F1) and XLM-R-base (79.2 F1). Surprisingly, the improvement is quite large ($+3.5$ to $+6.3$ F1) for seven out of ten African languages: yor, luo, lug, kin, ibo, and amh. The most interesting observation is that AfroXLMR-large, on average, is either competitive or better than the individual language LAFT models, including languages not seen during the MAFT training stage like lug, luo and wol. This implies that AfroXLMR-large (a single model) provides a better alternative to XLM-R-large+LAFT (for each language) in terms of performance on downstream tasks and disk space. AfroXLMR-large is currently the largest masked language model for African languages, and achieves the state-of-the-art compared to all other multilingual PLM on the NER task. This shows that our MAFT approach is very effective and scales to larger PLMs.

6 Cross-lingual Transfer Learning

The previous section demonstrates the applicability of MAFT in the fully-supervised transfer learning setting. Here, we demonstrate that our MAFT approach is also very effective in the zero-shot cross-lingual transfer setting using parameter-efficient fine-tuning methods.

Parameter-efficient fine-tuning methods like *adapters* (Houlsby et al., 2019) are appealing because of their modularity, portability, and composability across languages and tasks. Often times, language adapters are trained on a general domain corpus

---

[11]For languages not in MasakhaNER, we use the same monolingual data in Table 9.

Model	Size	amh	ara	eng	hau	ibo	kin	lug	luo	pcm	swa	wol	xho	yor	avg
XLM-R-large	550M	76.2	79.7	**93.1**	90.5	84.1	73.8	81.6	73.6	89.0	89.4	67.9	72.4	78.9	80.8
XLM-R-large+LAFT	550M x 13	**79.9**	**81.3**	92.2	**91.7**	**87.7**	78.4	86.2	**78.2**	**91.1**	90.3	68.8	72.7	82.9	83.2
AfroXLMR-large	550M	79.7	80.9	92.2	91.2	**87.7**	**79.1**	**86.7**	78.1	91.0	**90.4**	**69.6**	**72.9**	**85.2**	**83.4**

Table 7: **NER** model comparison on XLM-R-large, XLM-R-large+LAFT and XLM-R-large+MAFT (i.e AfroXLMR-large), showing F1-score on the test sets after 50 epochs averaged over 5 runs. Results are for all 4 tags in the dataset: PER, ORG, LOC, DATE/MISC.

Model	amh	hau	ibo	kin	lug	luo	pcm	swa	wol	yor	avg
XLM-R-base (fully-supervised)	69.7	91.0	86.2	**73.8**	80.5	75.8	**86.9**	**88.7**	69.6	78.1	81.2
mBERT (MAD-X) (Ansell et al., 2021)	-	83.4	71.7	65.3	67.0	52.2	72.1	77.6	65.6	74.0	69.9
mBERT (MAD-X on **news domain**)	-	86.0	77.6	69.9	73.3	56.9	78.5	80.2	68.8	75.6	74.1
XLM-R-base (MAD-X on **news domain**)	47.5	85.5	83.2	72.0	75.7	57.8	76.8	84.0	68.2	72.2	75.0
AfroXLMR-base (MAD-X on **news domain**)	47.7	88.1	80.9	73.0	80.1	59.2	79.9	86.9	69.1	75.6	77.0
mBERT (LT-SFT) (Ansell et al., 2021)	-	83.5	76.7	67.4	67.9	54.7	74.6	79.4	66.3	74.8	71.7
mBERT (LT-SFT on **news domain**)	-	86.4	80.6	69.2	76.8	55.1	80.4	82.3	71.6	76.7	75.4
XLM-R-base (LT-SFT on **news domain**)	<u>54.1</u>	87.6	81.4	72.7	79.5	<u>60.7</u>	81.2	85.5	73.6	73.7	77.3
AfroXLMR-base (LT-SFT on **news domain**)	54.0	<u>88.6</u>	<u>83.5</u>	<u>73.8</u>	<u>81.0</u>	<u>60.7</u>	<u>81.7</u>	86.4	<u>74.5</u>	<u>78.7</u>	<u>78.8</u>

Table 8: **Cross-lingual transfer using LT-SFT** (Ansell et al., 2021) and evaluation on **MasakhaNER**. The full-supervised baselines are obtained from Adelani et al. (2021) to measure performance gap when annotated datasets are available. Experiments are performed on 3 tags: PER, ORG, LOC. Average (avg) excludes amh. The best zero-shot transfer F1-scores are underlined.

like Wikipedia. However, when there is a mismatch between the target *domain* of the task and the *domain* of the language adapter, it could also impact the cross-lingual performance.

Here, we investigate how we can improve the cross-lingual transfer abilities of our adapted PLM – AfroXLMR-base by training language adapters on the same domain as the target task. For our experiments, we use the MasakhaNER dataset, which is based on the news domain. We compare the performance of language adapters trained on *Wikipedia* and *news* domains. In addition to adapters, we experiment with another parameter-efficient method based on Lottery-Ticket Hypothesis (Frankle & Carbin, 2019) i.e. LT-SFT (Ansell et al., 2021).

For the adapter approach, we make use of the MAD-X approach (Pfeiffer et al., 2020) – an adapter-based framework that enables cross-lingual transfer to arbitrary languages by learning modular language and task representations. However, the evaluation data in the target languages should have the same task and label configuration as the source language. Specifically, we make use of MAD-X 2.0 (Pfeiffer et al., 2021) where the last adapter layers are dropped, which has been shown to improve performance. The setup is as follows: (1) We train language adapters via masked language modelling (MLM) individually on source and target languages, the corpora used are described in Appendix A.2; (2) We train a task adapter by fine-tuning on the target task using labelled data in a source language. (3) During inference, task and language adapters are stacked together by substituting the source language adapter with a target language adapter.

We also make use of the Lottery Ticket Sparse

Fine-tuning (LT-SFT) approach (Ansell et al., 2021), a parameter-efficient fine-tuning approach that has been shown to give competitive or better performance than the MAD-X 2.0 approach. The LT-SFT approach is based on the Lottery Ticket Hypothesis (LTH) that states that each neural model contains a sub-network (a "winning ticket") that, if trained again in isolation, can reach or even surpass the performance of the original model. The LTH is originally a compression approach, the authors of LT-SFT re-purposed the approach for cross-lingual adaptation by finding sparse sub-networks for tasks and languages, that will later be composed together for zero-shot adaptation, similar to Adapters. For additonal details we refer to Ansell et al. (2021).

6.1 Experimental setup

For our experiments, we followed the same setting as Ansell et al. (2021) that adapted mBERT from English CoNLL03 (Tjong Kim Sang & De Meulder, 2003) to African languages (using MasakhaNER dataset) for the NER task.[12] Furthermore, we extend the experiments to XLMR-base and AfroXLMR-base. For the training of MAD-X 2.0 and sparse fine-tunings (SFT) for African languages, we make use of the monolingual texts from the news domain since it matches the domain of the evaluation data. Unlike, Ansell et al. (2021) that trained adapters and SFT on monolingual data from Wikipedia domain except for luo and pcm where the dataset is absent, we show that the domain used for training language SFT is also very important. For a

---

[12]We excluded the MISC and DATE from CoNLL03 and MasakhaNER respectively to ensure same label configuration.

fair comparison, we reproduced the result of Ansell et al. (2021) by training MAD-X 2.0 and LT-SFT on mBERT, XLM-R-base and AfroXLMR-base on target languages with the news domain corpus. But, we still make use of the pre-trained English language adapter[13] and SFT[14] for mBERT and XLM-R-base trained on the Wikipedia domain. For the AfroXLMR-base, we make use of the same English adapter and SFT as XLM-R-base because the PLM is already good for English language. We make use of the same hyper-parameters reported in the LT-SFT paper.

Hyper-parameters for adapters We train the task adapter using the following hyper-parameters: batch size of 8, 10 epochs, "pfeiffer" adapter config, adapter reduction factor of 8, and learning rate of 5e-5. For the language adapters, we make use of 100 epochs or maximum steps of 100K, minimum number of steps is 30K, batch size of 8, "pfeiffer+inv" adapter config, adapter reduction factor of 2, learning rate of 5e-5, and maximum sequence length of 256. For a fair comparison with adapter models trained on Wikipedia domain, we used the same hyper-parameter settings (Ansell et al., 2021) for the news domain.

6.2 Results and discussion

Table 8 shows the results of MAD-X 2.0 and LT-SFT, we compare their performance to fully supervised setting, where we fine-tune XLM-R-base on the training dataset of each of the languages, and evaluate on the test-set. We find that both MAD-X 2.0 and LT-SFT using news domain for African languages produce better performance (+4.2 on MAD-X and +3.7 on LT-SFT) than the ones trained largely on the wikipedia domain. This shows that the domain of the data matters. Also, we find that training LT-SFT on XLM-R-base gives better performance than mBERT on all languages. For MAD-X, there are a few exceptions like `hau`, `pcm`, and `yor`. Overall, the best performance is obtained by training LT-SFT on AfroXLMR-base, and sometimes it give better performance than the fully-supervised setting (e.g. as observed in `kin` and `lug`, `wol` `yor` languages). On both MAD-X and LT-SFT, AfroXLMR-base gives the best result since it has been firstly adapted on several African languages and secondly on the target domain of the target task. This shows that the MAFT approach is effective since the technique provides a better PLM that parameter-efficient methods can benefit from.

7 Conclusion

In this work, we proposed and studied MAFT as an approach to adapt multilingual PLMs to many African languages with a single model. We evaluated our approach on 3 different NLP downstream tasks and additionally contribute novel news topic classification dataset for 4 African languages. Our results show that MAFT is competitive to LAFT while providing a single model compared to many models specialized for individual languages. We went further to show that combining vocabulary reduction and MAFT leads to a 50% reduction in the parameter size of a XLM-R while still being competitive to applying LAFT on individual languages. We hope that future work improves vocabulary reduction to provide even smaller models with strong performance on distant and low-resource languages. To further research on NLP for African languages and reproducibility, we have uploaded our language adapters, language SFTs, AfroXLMR-base, AfroXLMR-small, and AfroXLMR-mini models to the HuggingFace Model Hub[15].

References

Amine Abdaoui, Camille Pradel, and Grégoire Sigel. Load what you need: Smaller versions of mutililingual BERT. In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pp. 119–123, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.sustainlp-1.16. URL https://aclanthology.org/2020.sustainlp-1.16.

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne

---

[13]https://adapterhub.ml/
[14]https://huggingface.co/cambridgeltl

[15]https://huggingface.co/models?sort=downloads&search=Davlan%2Fafro-xlmr

Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3053–3070, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main. 223. URL https://aclanthology.org/2022.naacl-main.223.

David Ifeoluwa 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, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131, 2021. doi: 10.1162/tacl_a_00416. URL https://aclanthology.org/2021.tacl-1.66.

Jesujoba Alabi, Kwabena Amponsah-Kaakyire, David Adelani, and Cristina España-Bonet. Massive vs. curated embeddings for low-resourced languages: the case of Yorùbá and Twi. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 2754–2762, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.335.

Alan Ansell, Edoardo Maria Ponti, Anna Korhonen, and Ivan Vulić. Composable sparse fine-tuning for cross-lingual transfer, 2021.

Israel Abebe Azime and Nebil Mohammed. An amharic news text classification dataset. *ArXiv*, abs/2103.05639, 2021.

Yassine Benajiba, Paolo Rosso, and José Miguel BenedíRuiz. Anersys: An arabic named entity recognition system based on maximum entropy. In Alexander Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing*, pp. 143–153, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-70939-8.

Michael Beukman. Analysing the effects of transfer learning on low-resourced named entity recognition performance. In *3rd Workshop on African Natural Language Processing*, 2021. URL https://openreview.net/forum?id=HKWMFqfN8b5.

Ethan C. Chau and Noah A. Smith. Specializing multilingual language models: An empirical study. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pp. 51–61, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.mrl-1.5. URL https://aclanthology.org/2021.mrl-1.5.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3576–3588, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main. 280. URL https://aclanthology.org/2021.naacl-main.280.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. Rethinking embedding coupling in pre-trained language models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=xpFFI_NtgpW.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard 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*, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main. 747. URL https://aclanthology.org/2020.acl-main.747.

Davis David. Swahili : News classification dataset. *Zenodo*, December 2020. doi: 10.5281/zenodo. 5514203. URL https://doi.org/10.5281/zenodo.5514203. The news version contains both train and test sets.

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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Roald Eiselen. Government domain named entity recognition for South African languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 3344–3348, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1533.

∀, Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online, 2020. URL https://www.aclweb.org/anthology/2020.findings-emnlp.195.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJl-b3RcF7.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8342–8360, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.740. URL https://aclanthology.org/2020.acl-main.740.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *ArXiv*, abs/2111.09543, 2021.

Michael A. Hedderich, David Adelani, Dawei Zhu, Jesujoba Alabi, Udia Markus, and Dietrich Klakow. Transfer learning and distant supervision for multilingual transformer models: A study on African languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2580–2591, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.204. URL https://aclanthology.org/2020.emnlp-main.204.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/houlsby19a.html.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4163–4174, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.372. URL https://aclanthology.org/2020.findings-emnlp.372.

Weijie Liu, Peng Zhou, Zhiruo Wang, Zhe Zhao, Haotang Deng, and Qi Ju. FastBERT: a self-distilling BERT with adaptive inference time. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6035–6044, Online, July 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.537. URL https://aclanthology.org/2020.acl-main.537.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020b. doi: 10.1162/tacl_a_00343. URL https://aclanthology.org/2020.tacl-1.47.

Fuli Luo, Wei Wang, Jiahao Liu, Yijia Liu, Bin Bi, Songfang Huang, Fei Huang, and Luo Si. VECO: Variable and flexible cross-lingual pre-training for language understanding and generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language*

Processing (Volume 1: Long Papers), pp. 3980–3994, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.308. URL https://aclanthology.org/2021.acl-long.308.

Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Sebastian Ruder, Ibrahim Said Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Abdullahi Salahudeen, Aremu Anuoluwapo, Alípio Jeorge, and Pavel Brazdil. Naijasenti: A nigerian twitter sentiment corpus for multilingual sentiment analysis. *ArXiv*, abs/2201.08277, 2022.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 448–462, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.38. URL https://aclanthology.org/2021.naacl-main.38.

Rubungo Andre Niyongabo, Qu Hong, Julia Kreutzer, and Li Huang. KINNEWS and KIRNEWS: Benchmarking cross-lingual text classification for Kinyarwanda and Kirundi. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 5507–5521, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.480. URL https://aclanthology.org/2020.coling-main.480.

Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. CAMeL tools: An open source python toolkit for Arabic natural language processing. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 7022–7032, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.868.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pp. 116–126, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.mrl-1.11. URL https://aclanthology.org/2021.mrl-1.11.

Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. ERNIE-M: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 27–38, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.3. URL https://aclanthology.org/2021.emnlp-main.3.

Chester Palen-Michel, June Kim, and Constantine Lignos. Multilingual open text 1.0: Public domain news in 44 languages. *CoRR*, abs/2201.05609, 2022. URL https://arxiv.org/abs/2201.05609.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 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)*, pp. 7654–7673, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.617. URL https://aclanthology.org/2020.emnlp-main.617.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. UNKs everywhere: Adapting multilingual language models to new scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10186–10203, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.800. URL https://aclanthology.org/2021.emnlp-main.800.

Sara Rosenthal, Noura Farra, and Preslav Nakov. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pp. 502–518, 2017.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019.

Kathleen Siminyu, Godson Kalipe, Davor Orlic, Jade Z. Abbott, Vukosi Marivate, Sackey Freshia, Prateek Sibal, Bhanu Neupane, David Ifeoluwa Adelani, Amelia Taylor, Jamiil Toure Ali, Kevin Degila, Momboladji Balogoun, Thierno Ibrahima Diop, Davis David, Chayma Fourati, Hatem Haddad, and Malek Naski. Ai4d - african language program. *ArXiv*, abs/2104.02516, 2021.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. MobileBERT: a compact task-agnostic BERT for resource-limited devices. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2158–2170, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.195. URL https://aclanthology.org/2020.acl-main.195.

Y. Tang, C. Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual translation with extensible multilingual pretraining and finetuning. *ArXiv*, abs/2008.00401, 2020.

Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pp. 142–147, 2003. URL https://aclanthology.org/W03-0419.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pretrained transformers. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 5776–5788. Curran Associates, Inc., 2020.

Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. MiniLMv2: Multi-head self-attention relation distillation for compressing pretrained transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 2140–2151, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.188. URL https://aclanthology.org/2021.findings-acl.188.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41.

Seid Muhie Yimam, Hizkiel Mitiku Alemayehu, Abinew Ayele, and Chris Biemann. 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), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.91. URL https://aclanthology.org/2020.coling-main.91.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *NIPS*, 2015.

A   Appendix

A.1   Monolingual corpora for LAFT and MAFT

For training the MAFT models, we make use of the aggregation of monolingual data from Table 9.

For the LAFT models, we make use of existing XLMR-base+LAFT models from the MasakhaNER paper (Adelani et al., 2021). However, for other languages not present in MasakhaNER (ara, mlg, orm, sna, som, xho), we make use of the mC4 corpus except for eng — we use the VOA corpus. For a fair comparison across models, when training the XLM-R-large+LAFT models, we use the same monolingual corpus used to train XLM-R-base+LAFT models.

A.2   News corpora for language adapters and SFTs

Table 10 provides the news corpus we used to train language adapters and SFTs for the cross-lingual settings.

Language	Source	Size (MB)	No. of sentences
Afrikaans (afr)	mC4 (subset) (Xue et al., 2021)	752.2MB	3,697,430
Amharic (amh)	mC4 (subset), and VOA	1,300MB	2,913,801
Arabic (ara)	mC4 (subset)	1,300MB	3,939,375
English (eng)	mC4 (subset), and VOA	2,200MB	8,626,571
French (fra)	mC4 (subset), and VOA	960MB	4,731,196
Hausa (hau)	mC4 (all), and VOA	594.1MB	3,290,382
Igbo (ibo)	mC4 (all), and AfriBERTa Corpus (Ogueji et al., 2021)	287.5MB	1,534,825
Malagasy (mlg)	mC4 (all)	639.6MB	3,304,459
Chichewa (nya)	mC4 (all), Chichewa News Corpus (Siminyu et al., 2021)	373.8MB	2,203,040
Oromo (orm)	AfriBERTa Corpus, and VOA	67.3MB	490,399
Naija (pcm)	AfriBERTa Corpus	54.8MB	166,842
Rwanda-Rundi (kin/run)	AfriBERTa Corpus, KINNEWS & KIRNEWS (Niyongabo et al., 2020), and VOA	84MB	303,838
chiShona (sna)	mC4 (all), and VOA	545.2MB	2,693,028
Somali (som)	mC4 (all), and VOA	1,000MB	3,480,960
Sesotho (sot)	mC4 (all)	234MB	1,107,565
Kiswahili (swa)	mC4 (all)	823.5MB	4,220,346
isiXhosa (xho)	mC4 (all), and Isolezwe Newspaper	178.4MB	832,954
Yorùbá (yor)	mC4 (all), Alaroye News, Asejere News, Awikonko News, BBC, and VON	179.3MB	897,299
isiZulu (zul)	mC4 (all), and Isolezwe Newspaper	700.7MB	3,252,035

Table 9: Monolingual Corpora (after pre-processing – we followed AfriBERTa (Ogueji et al., 2021) approach) , their sources and size (MB), and number of sentences.

Language	Source	Size (MB)	No. of sentences
Amharic (amh)	VOA (Palen-Michel et al., 2022)	19.9MB	72,125
Hausa (hau)	VOA (Palen-Michel et al., 2022)	46.1MB	235,614
Igbo (ibo)	BBC Igbo (Ogueji et al., 2021)	16.6MB	62,654
Kinyarwanda (kin)	KINNEWS (Niyongabo et al., 2020)	35.8MB	61,910
Luganda (lug)	Bukedde	7.9MB	67,716
Luo (luo)	Ramogi FM news and MAFAND-MT (Adelani et al., 2022)	1.4MB	8,684
Naija (pcm)	BBC	50.2MB	161,843
Kiswahili (swa)	VOA (Palen-Michel et al., 2022)	17.1MB	88,314
Wolof (wol)	Lu Defu Waxu, Saabal, Wolof Online, and MAFAND-MT (Adelani et al., 2022)	2.3MB	13,868
Yorùbá (yor)	BBC Yorùbá	15.0MB	117,124

Table 10: Monolingual News Corpora used for language adapter and SFT training, their sources and size (MB), and number of sentences.