Small Data? No Problem! Exploring the Viability of Pretrained Multilingual Language Models for Low-Resource Languages

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin

David R. Cheriton School of Computer Science
University of Waterloo

{kelechi.ogueji, yuxin.zhu, jimmylin}@uwaterloo.ca

Abstract

Pretrained multilingual language models have been shown to work well on many languages for a variety of downstream NLP tasks. However, these models are known to require a lot of training data. This consequently leaves out a huge percentage of the world's languages as they are under-resourced. Furthermore, a major motivation behind these models is that lower-resource languages benefit from joint training with higher-resource languages. In this work, we challenge this assumption and present the first attempt at training a multilingual language model on only low-resource languages. We show that it is possible to train competitive multilingual language models on less than 1 GB of text. Our model, named AfriBERTa, covers 11 African languages, including the first language model for 4 of these languages. Evaluations on named entity recognition and text classification spanning 10 languages show that our model outperforms mBERT and XLM-R in several languages and is very competitive overall. Results suggest that our "small data" approach based on similar languages may sometimes work better than joint training on large datasets with high-resource languages. Code, data and models are released at https://github.com/keleog/afriberta.

1 Introduction

Pretrained language models have risen to the fore of natural language processing (NLP), achieving impressive performance on a variety of NLP tasks. The multilingual version of these models such as XLM-R (Conneau et al., 2020) and mBERT (Devlin et al., 2019) have also been shown to generalize well to many languages. However, these models are known to require a lot of training data, which is often absent for low-resource languages. Also, high-resource languages usually make up a significant part of the training data, as it is hypothesized that they help boost transfer to lower-resource languages. Hence, there has been no known attempt to investigate if it is possible to pretrain multilingual language models solely on low-resource languages without any transfer from higher-resource languages, despite the numerous benefits that this could provide. Motivated by this gap in the literature, the goal of our work is to explore the viability of multilingual language models pretrained from scratch on low-resource languages and to understand how to pretrain such models in this setting.

We introduce AfriBERTa, a transformer-based multilingual language models trained on 11 African languages, all of which are low-resource.[1] We evaluate this model on named entity recognition (NER) and text classification downstream tasks on 10 low-resource languages. Our models outperform larger models like mBERT and XLM-R by up to 10 F1 points on text classification, and also outperform these models on several languages in the NER task. Across all languages, we obtain very competitive performance to these larger models. In summary, our contributions are as follows:

1. We show that competitive multilingual language models can be pretrained from scratch solely on low-resource languages without any high-resource transfer.

2. We show that it is possible to pretrain these models on less than 1 GB of text data and highlight the many practical benefits of this.

3. Our extensive experiments highlight important factors to consider when pretraining multilingual language models in low-resource settings.

4. We introduce language models for 4 languages, improving the representation of low-resource languages in modern NLP tools.

[1] One of the languages (Gahuza) is counted twice because it is a code-mixed language consisting of Kinyarwanda and Kirundi.

Our results show that, for the first time, it is possible to pretrain a multilingual language model from scratch on only low-resource languages and obtain good performance on downstream tasks.

2 Related Work

In recent years, unsupervised learning of text representations has significantly advanced natural language processing tasks. Static representations from pretrained word embeddings (Mikolov et al., 2013; Pennington et al., 2014) were improved upon by learning contextualized representations (Peters et al., 2018). This has been noticeably improved further by pretraining language models (Radford et al., 2018; Devlin et al., 2019) based on transformers (Vaswani et al., 2017). These models have also been extended to the multilingual setting where a single language model is pretrained on several languages without any explicit cross-lingual supervision (Conneau et al., 2020; Devlin et al., 2019).

However, much of this progress has been focused on languages with relatively large amounts of data, commonly referred to as *high-resource languages*. There has especially been very little focus on African languages, despite the over 2000 languages spoken on the continent making up 30.1% of all living languages (Eberhard et al., 2019). This is further visible in NLP publications on these languages. In all the Association for Computational Linguistics (ACL) conferences hosted in 2019, only 0.19% author affiliations were located in Africa (Caines, 2019). Other works (Joshi et al., 2020) have also noted the great disparity in the coverage of languages by NLP technologies. They note that over 90% of the world's 7000+ languages are under-studied by the NLP community.

There have been a few works on learning pretrained embeddings for African languages, although many of them have been static and trained on a specific language (Ezeani et al., 2018; Ogueji and Ahia, 2019; Alabi et al., 2019; Dossou and Sabry, 2021). More recently, Azunre et al. (2021) trained a BERT model on the Twi language. However, they note that their model is biased to the religious domain because much of their data comes from that domain.

While some African languages have been included in multilingual language models, this coverage only scratches the surface of the number of spoken African languages. Furthermore, the languages always make up a minuscule percentage of the training set. For instance, amongst the 104 languages that mBERT was pretrained on, only 3 are African.[2] In XLM-R, there are only 8 African languages out of the 100 languages. In terms of dataset size, the story is the same. African languages make up 4.80 GB out of about 2395 GB that XLM-R was pretrained on, representing just 0.2% of the entire dataset (Conneau et al., 2020). In mBERT, African languages make up just 0.24 GB out of the approximately 100 GB that the model was pretrained on. All of this call for an obvious need for increased representation of African languages in modern NLP tools for the over 1.3 billion speakers on the continent.[3]

Pretrained language models have been shown to perform well when there is a lot of data (Liu et al., 2019; Conneau et al., 2020), but some works have focused on using relatively smaller amounts of data. Martin et al. (2020) showed that it is possible to obtain state-of-the-art result with a French BERT model pretrained on small-scale diverse data. In another work, Micheli et al. (2020) showed that training a French BERT language model on 100 MB of data yields similar performance on question answering as models pretrained on larger datasets. Furthermore, Ortiz Suárez et al. (2020) obtained state-of-the-art performance with ELMo (Peters et al., 2018) language models pretrained on less than 1 GB of Wikipedia text, and Zhang et al. (2020) show that RoBERTa language models (Liu et al., 2019) trained on 10 to 100 million tokens can encode most syntactic and semantic features in its learned text representations.

A common theme among these works is their focus on monolingual language models. While it is possible to learn monolingual language models on smaller amounts of data, it remains to be seen if it is possible in the multilingual case. Our work is the first, to the best of our knowledge, that focuses on pretraining a multilingual language model solely on low-resource languages without any transfer from higher-resource languages.

3 Methodology

3.1 Data

Languages: We focus on 11 African languages, namely Afaan Oromoo (also called Oromo),

[2] https://github.com/google-research/bert/blob/master/multilingual.md

[3] https://www.worldometers.info/world-population/africa-population/ (accessed on February 19, 2021)

Language	Family	Speakers	Region
Afaan Oromoo	Afro-Asiatic	50M	East
Amharic	Afro-Asiatic	26M	East
Gahuza	Niger-Congo	21M	East
Hausa	Afro-Asiatic	63M	West
Igbo	Niger-Congo	27M	West
Nigerian Pidgin	English Creole	75M	West
Somali	Afro-Asiatic	19M	East
Swahili	Niger-Congo	98M	Central/East
Tigrinya	Afro-Asiatic	7M	East
Yorùbá	Niger-Congo	42M	West

Table 1: **Language Information:** For each language, its family, number of speakers (Eberhard et al., 2019), and regions in Africa spoken.

Language	# Sent.	# Tok.	Size (GB)
Afaan Oromoo	410,840	6,870,959	0.051
Amharic	525,024	1,303,086	0.213
Gahuza	131,952	3,669,538	0.026
Hausa	1,282,996	27,889,299	0.150
Igbo	337,081	6,853,500	0.042
Nigerian Pidgin	161,842	8,709,498	0.048
Somali	995,043	27,332,348	0.170
Swahili	1,442,911	30,053,834	0.185
Tigrinya	12,075	280,397	0.027
Yorùbá	149,147	4,385,797	0.027
Total	5,448,911	108,800,600	0.939

Table 2: **Dataset Size:** Size of each language in the dataset covering numbers of sentences, tokens and uncompressed disk size.

Amharic, Gahuza (a code-mixed language containing Kinyarwanda and Kirundi), Hausa, Igbo, Nigerian Pidgin, Somali, Swahili, Tigrinya and Yorùbá. These languages all come from three language families: Niger-Congo, Afro Asiatic and English Creole. We select these languages because they are the languages supported by the British Broadcasting Corporation (BBC) News, which was our main source of data.[4] We also obtain additional data from the Common Crawl Corpus (Conneau et al., 2020; Wenzek et al., 2020) for languages available there, specifically Amharic, Afaan Oromoo, Amharic, Hausa, Igbo, Somali and Swahili. Table 1 provides details about the languages used in pretraining our models.

Size: The total size of our pretraining corpus is 0.94 GB (108.8 million tokens). In comparison, XLM-R was pretrained on about 2395 GB (164.0 billion tokens) (Conneau et al., 2020), and mBERT was trained on roughly 100 GB (12.8 billion tokens).[5] Following findings from Liu et al. (2019) and Conneau et al. (2020) that more data is always better for pretrained language modelling, our small corpus makes our task even more challenging, and one can already see that our model is at a disadvantage compared to XLM-R and mBERT.

Language	XLM-R	mBERT	AfriBERTa
Afaan Oromoo	0.10	-	0.05
Amharic	0.80	-	0.21
Hausa	0.30	-	0.15
Somali	0.40	-	0.17
Swahili	1.60	0.04	0.19
Yorùbá	-	0.06	0.03

Table 3: **Comparing Sizes Across Models:** Comparison of the dataset sizes (GB) of languages present in XLM-R, mBERT and AfriBERTa. "-" indicates language was not present in model's pretraining corpus.

Our corpus contains approximately 5.45 million sentences and 108.8 million tokens. Table 2 presents more details about the dataset size for each language. It can be observed that languages like Swahili, Hausa and Somali have the most amount of data, while languages like Tigrinya have very little data with just about 12,000 sentences.

For each language we pretrained on that is present in XLM-R or mBERT, we compare the size of that language in our dataset to its size in the pretraining corpora of mBERT and XLM-R. From the comparison details in Table 3, we can see that XLM-R always has more data for languages present in our pretraining corpus and theirs. In fact, on average, we can see that the size of the language is always at least two times more in XLM-R. For mBERT, we can see that AfriBERTa has more data for Hausa and Yorùbá, which are present in both corpora. However, one would expect that, given that both languages are in the Latin script, there should be enough high-resource transfer to help them outperform our model.

Preprocessing: We remove lines that are empty or only contain punctuation. Given that there is significant overlap between the African language corpora in Common Crawl and the BBC News data that we crawled, we perform extensive deduplication for each language by removing exact matched sentences. We also enforce a minimum length restriction by only retaining sentences with more than 5 tokens. We observe that the quality of the dataset

[4] https://www.bbc.co.uk/ws/languages (scraped up to January 17, 2021)

[5] https://github.com/mayhewsw/multilingual-data-stats/tree/main/wiki

from Common Crawl is very low, confirming recent findings from Caswell et al. (2021). Hence, we manually clean the data as much as we can by removing texts in the wrong language, while trying to throw out as little data as possible.

3.2 Model

We train a transformer (Vaswani et al., 2017) with the standard masked language modelling objective of Devlin et al. (2019) without next sentence prediction. This is also the same approach used in XLM-R (Conneau et al., 2020). We pretrain on text data containing all languages, sampling batches from different languages. We sample languages such that our model does not see the same language over several consecutive batches.

We utilize subword tokenization on the raw text data using SentencePiece (Kudo and Richardson, 2018) trained with a unigram language model (Kudo, 2018). We sample training sentences from different languages for the tokenizer following the sampling method described in Conneau and Lample (2019) with $\alpha = 0.3$.

3.3 Evaluation

Pretraining: We take out varying amounts of evaluation sentences from each language's original monolingual dataset, depending on the language's size. Our total evaluation set containing all languages consists of roughly 440,000 sentences. We evaluate the perplexity on this dataset to measure language model performance. However, following Conneau et al. (2020), we continue pretraining even after validation perplexity stops decreasing. Effectively, we pretrain on around 0.94 GB of data and evaluate on around 0.08 GB of data.

NER: We evaluate named entity recognition (NER) using the recently released MasakhaNER dataset (Adelani et al., 2021). The dataset covers the following ten languages: Amharic, Hausa, Igbo, Kinyarwanda, Luganda, Luo, Nigerian Pidgin, Swahili, Wolof and Yorùbá. The authors established strong baselines on the dataset ranging from simpler methods like CNN-BiLSTM-CRF to pretrained language models like mBERT and XLM-R.

Text Classification: We use the news topic classification dataset from Hedderich et al. (2020), which covers Hausa and Yoruba. The authors established strong transfer learning and distant supervision baselines. They find that both mBERT and XLM-R outperform simpler neural network baselines in few-shot and zero-shot settings.

3.4 Experimental Setup

All models are trained with the Huggingface Transformers library (Wolf et al., 2020) (v4.2.1). In the following initial experiments, we pretrain each model for 60,000 steps and use a maximum sequence length of 512. We pretrain using a batch size of 32 and accumulate the gradients for 4 steps. Optimization is done using AdamW (Loshchilov and Hutter, 2017) with a learning rate of 1e-4 and 6000 linear warm-up steps. We report F1 scores on the NER dataset averaged over 3 runs with different random seeds. Following initial explorations, we found a vocabulary size of 40k, excluding special tokens, to yield good results across different model sizes, so we use this for initial experiments.

NER models are trained by adding a linear classification layer to the pretrained transformer model and fine-tuning all parameters. Following Adelani et al. (2021), we train for 50 epochs with a batch size of 16, a learning rate of 5e-5 and also optimize with AdamW.

Text classification models are trained by adding a linear classification layer to the pretrained transformer model and fine-tuning all parameters. We train for 25 epochs with a batch size of 32, warm-up steps of 100, learning rate of 5e-5 and optimize with AdamW as well.

4 Results

4.1 Design Space Exploration

In this section, we compare variants of AfriBERTa models to each other in a bid to understand how to pretrain multilingual language models in small data regimes. We pretrain variants from the point of view of model architecture, taking three factors into consideration: (i) model depth, (ii) number of attention heads and (iii) vocabulary size. We define performance as "good transfer to downstream task". Because the NER dataset covers more languages, we fine-tune and evaluate our models on it.

Model Depth: We compare models with 4, 6, 8 and 10 layers. For each model, we use 4 attention heads and adjust the size of the hidden units and feed-forward layers so that all models have approximately the same number of parameters. From preliminary experiments, models with more than 10 layers did not yield substantially better performance. This is expected, given the small size of the

# Layers	# Params	amh	hau	ibo	kin	lug	luo	pcm	swa	wol	yor	avg
4	74.8M	62.18	89.66	87.03	69.29	67.23	59.00	83.57	83.89	77.04	67.02	75.97
6	74.7M	61.59	90.34	85.81	72.76	66.39	61.43	86.27	84.02	76.61	68.54	76.91
8	74.6M	62.04	90.96	86.33	74.00	68.66	60.96	84.43	84.16	76.11	67.38	77.00
10	74.3M	62.14	90.69	87.36	75.74	67.87	60.59	84.79	84.70	76.17	67.51	**77.27**

Table 4: **Effect of Number of Layers:** NER dev F1 scores (averaged over three different random seeds) on each language for models with different layer depth, but same number of parameters. The sizes of the embedding and feed-foward layers are adjusted such that feed-foward is always approximately 4 times embedding size. The highest F1-score per language is underlined, while the highest overall average is in **bold**.

# Layers	# Att. Heads	# Params	amh	hau	ibo	kin	lug	luo	pcm	swa	wol	yor	avg
4	2	60.1M	58.23	88.78	84.63	71.28	65.68	56.91	83.84	82.44	76.69	64.64	74.99
4	4	60.1M	60.09	89.34	87.08	72.95	68.25	60.10	84.08	83.17	76.29	66.73	76.44
4	6	60.1M	60.26	89.49	86.01	72.69	67.82	59.85	84.68	83.73	76.22	67.66	76.46
6	2	74.3M	60.54	89.72	87.25	72.68	70.23	59.98	84.52	83.25	76.00	67.00	76.74
6	4	74.3M	63.29	90.19	86.05	74.26	68.58	59.23	84.74	83.46	77.62	67.04	76.80
6	6	74.3M	60.38	90.86	86.70	73.12	68.54	61.68	84.59	82.80	79.02	68.48	77.31
8	2	88.5M	60.32	90.55	85.32	75.38	69.89	62.73	85.50	83.51	79.07	68.09	77.78
8	4	88.5M	61.90	90.79	86.67	74.28	68.45	61.57	85.64	83.88	78.48	70.16	77.77
8	6	88.5M	60.92	90.16	86.95	74.71	70.66	60.75	85.48	84.87	78.04	71.16	**78.09**
10	2	102.6M	59.87	90.78	87.10	73.73	66.29	60.03	85.04	83.47	81.12	69.06	77.40
10	4	102.6M	63.95	91.33	87.11	75.24	68.96	63.36	85.66	84.67	74.60	69.27	77.80
10	6	102.6M	63.94	90.54	87.39	75.90	69.19	61.73	85.77	84.66	75.64	69.48	77.81

Table 5: **Effect of Number of Attention Heads:** NER dev F1 scores (averaged over three different random seeds) on each language for different models with the same number of layers, but different number of attention heads. The highest F1-score per layer size is underlined, while the highest overall average is in **bold**.

data. Because of this, coupled with computational constraints, we do not explore settings with more than 10 layers.

As we can see from the results in Table 4, deeper models always outperform shallower models. However, the performance gain diminishes with size. For example, the gain from increasing the model to 6 layers from 4 layers is roughly 1 F1 point. However, the gain from increasing from 6 layers to 10 layers is only $\sim$0.4. This corroborates the recent *universality overfitting* findings from Kaplan et al. (2020), who showed that the performance of transformer language models improves predictably as long as data size and model depth are scaled in tandem, otherwise there is a diminishing return.

In general, our results suggests that deeper models also work well when pretraining multilingual language models on small datasets. This follows previous works on understanding the cross-lingual ability of multilingual language models (K et al., 2019), which have shown that deeper models have better cross-lingual performance. However, gains from increasing depth are relatively minimal because of the size of our corpus.

Number of Attention Heads: For each layer size (4, 6, 8 and 10), we train models with three different numbers of attention heads: 2, 4 and 6. Again, initial experiments with more than 6 attention heads did not yield any better results, so we do not explore more than 6 heads. Results are presented in Table 5.

The results suggest that there is a diminishing return to the number of attention heads when the model is deep. Shallower models need more attention heads to attain competitive performance. However, when the model is deep enough, it is very competitive with as few as two attention heads. This suggests that results from recent works (K et al., 2019; Michel et al., 2019), which suggest that transformers can do without a large number of attention heads, also hold true for multilingual language models on small datasets.

Vocabulary Size: Previous works have suggested that on small datasets, one should employ a small vocabulary size (Sennrich and Zhang, 2019; Araabi and Monz, 2020). However, it remains to be seen if this holds in the multilingual setting since several languages will be competing for vocabulary

# Layers	# Att. Heads	Vocab Size	# Params	amh	hau	ibo	kin	lug	luo	pcm	swa	wol	yor	avg
8	6	25k	76.9M	60.56	89.96	85.84	73.23	69.67	61.86	85.11	84.34	75.40	68.35	77.09
8	6	40k	88.5M	60.92	90.16	86.95	74.71	70.66	60.75	85.48	84.87	78.04	71.16	78.09
8	6	55k	99.9M	63.65	90.17	87.28	72.47	67.47	61.49	85.59	85.09	77.56	69.06	77.35
8	6	70k	111.5M	66.17	91.25	87.74	77.44	68.29	59.91	87.00	87.05	77.49	68.82	**78.33**
8	6	85k	123.1M	62.35	90.42	87.44	77.01	68.20	61.98	86.46	85.87	72.84	70.14	77.82

Table 6: **Effect of Vocabulary Size:** NER dev F1 scores (averaged over three different random seeds) on the best model size with varying vocabulary sizes. The highest overall average F1-score is in **bold**.

Language	In mBERT	In XLM-R?	In AfriBERTa?	CNN-BiLSTM CRF	mBERT (172M)	XLM-R base (270M)	AfriBERTa small (97M)	AfriBERTa base (111M)	AfriBERTa large (126M)
amh	no	yes	yes	52.89	0.0	70.96	67.90	71.80	**73.82**
hau	no	yes	yes	83.70	87.34	89.44	89.01	90.10	**90.17**
ibo	no	no	yes	78.48	85.11	84.51	86.63	86.70	**87.38**
kin	no	no	yes	64.61	70.98	**73.93**	69.91	73.22	73.78
lug	no	no	no	74.31	80.56	**80.71**	76.44	79.30	78.85
luo	no	no	no	66.42	72.65	**75.14**	67.31	70.63	70.23
pcm	no	no	yes	66.43	**87.78**	87.39	82.92	84.87	85.70
swa	yes	yes	yes	79.26	86.37	87.55	85.68	**88.00**	87.96
wol	no	no	no	60.43	**66.10**	64.38	60.10	61.82	61.81
yor	yes	no	yes	67.07	78.64	77.58	76.08	79.36	**81.32**
avg	–	–	–	69.36	71.55	**79.16**	76.20	78.60	79.10
avg (excl. amh)	–	–	–	71.19	79.50	**80.07**	77.12	79.36	79.69

Table 7: **Comparison of NER Results:** F1-scores on the test sets of each language. XLM-R and mBERT results obtained from Adelani et al. (2021). The best score for each language and overall best scores are in **bold.** We also report the model parameter size in parentheses.

space and Conneau et al. (2020) have found that increasing the vocabulary size improves multilingual performance. We evaluate our best model size on varying vocabulary sizes and report results in Table 6. As we can see from the results, increasing the vocabulary size does not always yield good results on smaller datasets. While a small vocabulary size performs relatively poorly, medium sized vocabularies can sometimes outperform larger ones. Due to computation constraints, we selected vocabulary size of 70k for the final models below.

Final Model Selection: We release three AfriBERTa pretrained model sizes: small (4 layers), base (8 layers) and large (10 layers). Each model has 6 attention heads, 768 hidden units, 3072 feed-forward size and a maximum length of 512. Their respective parameter sizes are 97 million, 111 million and 126 million. We use float16 operations to speed up training and reduce memory usage. Pretraining is done for 460,000 steps with 40,000 linear warm-up steps and then the learning rate is decreased linearly. We pretrain with a batch size of 32 on 2 Nvidia V100 GPUs and accumulate the gradients for 8 steps.

4.2 NER Comparisons

As we can see in Table 7, even the AfriBERTa small model, which is almost three times smaller than XLM-R, obtains competitive NER results across all languages, trailing XLM-R by less than 3 F1 points. This represents a great opportunity for deployment in resource constrained scenarios, which is usually common for applications in low-resource languages. Our best performing model is AfriBERTa large, which outperforms mBERT and is very competitive with XLM-R across all languages. AfriBERTa large even outperforms both models on several languages that all three models were pretrained on, such as Hausa, Amharic and Swahili.

It should be noted that AfriBERTa large achieves all this with less than half of the number of parameters of XLM-R and about 45M fewer parameters than mBERT. Furthermore, we can see that our models performs very well on languages that were not part of our pretraining corpus, such as Luo, Wolof and Luganda. This demonstrates its strong cross-lingual capabilities, despite smaller parameter sizes and pretraining corpus size. A notable observation is that both mBERT and XLM-R out-

Language	In mBERT	In XLM-R?	In AfriBERTa?	mBERT	XLM-R base	AfriBERTa large
hau	no	yes	yes	83.03	85.62	**90.86**
yor	yes	no	yes	71.61	71.07	**83.22**

Table 8: **Comparison of Text Classification Results:** F1-scores on the test sets. The best score for each language is in **bold**.

perform AfriBERTa on Nigerian Pidgin, despite not being trained on the language. This is likely because of the language's high similarity with English. Nigerian Pidgin is an English Creole, meaning it borrows and shares a lot of its properties (including words) with English. Since both mBERT and XLM-R were pretrained on very large amounts of English data, it is no surprise that they perform so well on Nigerian Pidgin. In summary, our small, base and large models' performance are comparable to mBERT and XLM-R across all languages, despite being pretrained on a substantially smaller corpus and having fewer model parameters.

4.3 Text Classification Comparisons

We also compare our best model (AfriBERTa large) to XLM-R base and mBERT on text classification. As we can see from the results in Table 8, AfriBERTa large clearly outperforms both XLM-R and mBERT by over 10 F1 points on Yorùbá and up to 7 F1 points on Hausa. Results show that mBERT slightly outperforms XLM-R on Yorùbá, most likely because it was pretrained on it, while XLM-R was not. XLM-R also outperforms mBERT on Hausa, presumably for the same reason. It should be noted that our model was pretrained on around half as much Hausa data as XLM-R, but still outperforms it substantially.

An important observation is that AfriBERTa outperforms both XLM-R and mBERT on text classification, but not so much on the NER task. This suggests that, perhaps, some downstream tasks benefit from larger multilingual models with high-resource transfer than other tasks. However, we leave this interesting observation for future work.

5 Discussion

In this section, we discuss some other contributions of this work. At a high level, AfriBERTa presents the first evidence that multilingual language models are viable with very little training data. This offers numerous benefits for the NLP community, especially for low-resource languages.

Opportunities for Smaller Curated Datasets: Our empirical results suggest that state-of-the-art NLP methods like multilingual language models can be made more accessible for low-resource languages. Caswell et al. (2021) recently showed that web-crawled multilingual corpora available for many languages, especially low-resource ones, are usually of very low quality. They found issues such as wrong-language content, erroneous language codes and low-quality sentences. Our work opens the door to competitive multilingual language models on smaller curated datasets for low-resource languages.

Another possible benefit of these smaller curated datasets is that they would tend to contain local content as opposed to foreign content as is in the Wikipedia and other relatively larger datasets of these languages. Models trained on such datasets with local content could potentially be more useful to the speakers of the languages given that they would be trained on data with local context.

Strength of Language Similarity: Our work challenges the commonly held belief in the NLP community that lower-resource languages need higher-resource languages in multilingual language models. Instead, we empirically demonstrate that pretraining on similar low-resource languages in a multilingual setting may sometimes be better than pretraining using high-resource and low-resource languages together. This approach should be considered in future work, especially since there have been recent findings (Wang et al., 2020) that low-resource languages also experience negative interference in multilingual models.

Potential Ethical Benefits: Recent works have called for more considerations of ethics and related concerns in the development of pretrained language models (Bender et al., 2021). These concerns have ranged from environmental and financial (Strubell et al., 2019) to societal bias (Kurita et al., 2019; Basta et al., 2019)

We believe our work offers the potential to ad-

Model	# Params	Data Size (GB)	# Tokens
XLM-R base	270M	2395	164.0B
mBERT	172M	100	12.8B
AfriBERTa base	112M	0.94	108.8M

Table 9: **Comparing Sizes:** Comparison of datasets and model sizes between XLM-R, mBERT and AfriBERTa.

dress some of these concerns, while developing language technology for under-served languages. A comparison of model and data sizes of common multilingual models is presented in Table 9. Smaller dataset sizes, like ours, mean that these datasets can more easily be cleaned, filtered, analyzed and *possibly* de-biased in comparison to the humongous data sizes of larger language models. We have also shown that smaller-sized models can outperform larger models, despite using smaller training resources. This represents a potential for reduced environmental impact.

While "low-resource" is commonly used in the NLP community to describe a lack of data resources, Nekoto et al. (2020) have argued that "low-resource" also includes a wide range of societal problems, including computational constraints. Thus, our work embodies the broader spirit of "low-resource", as we develop more efficient models on smaller data sizes for under-served languages.

Improving the Representation of African Languages in Modern NLP tools: As discussed in section 2, there is very poor representation of African languages in modern NLP tools. Recently, there have been significant efforts towards closing this gap (Alabi et al., 2019; Ogueji and Ahia, 2019; Nekoto et al., 2020; Ahia and Ogueji, 2020; Fan et al., 2020; Azunre et al., 2021; Dossou and Sabry, 2021; Adelani et al., 2021). Our work follows along this path, as there is a need to build language technologies for the over 1.3 billion people on the continent. Besides showing that multilingual language models are viable on low-resource African languages with small training data, we also introduce the first language models for four of these languages: Kinyarwanda, Kirundi, Nigerian Pidgin and Tigrinya. These are four languages with over 50 million speakers (Eberhard et al., 2019) who are active users of digital tools. However, these languages have noticeably deficient support in NLP technologies. Our work represents an important step towards improving this.

6 Conclusion

In this work, we introduced AfriBERTa, a multilingual language model pretrained on less than 1 GB of data from 11 African languages. We show that this model is competitive with models pretrained on larger datasets and even outperforms them on some languages. Our comprehensive experiments also highlight important factors to consider when pretraining multilingual language models on smaller datasets. More importantly, we highlight some practical benefits of viable language models on smaller datasets. Finally, we release code, pretrained models and the dataset to stimulate further work on multilingual language models for low-resource languages.

Acknowledgements

This research was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada and an AI for Social Good grant from the Waterloo AI Institute; computational resources were provided by Compute Ontario and Compute Canada.

References

David Ifeoluwa Adelani, Jade Z. 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 Hassan Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba O. Alabi, Seid Muhie Yimam, Tajuddeen Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin P. 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. 2021. MasakhaNER: Named entity recognition for African languages. *CoRR*, abs/2103.11811.

Orevaoghene Ahia and Kelechi Ogueji. 2020. Towards supervised and unsupervised neural machine

translation baselines for Nigerian Pidgin. *CoRR*, abs/2003.12660.

Jesujoba O. Alabi, Kwabena Amponsah-Kaakyire, David Ifeoluwa Adelani, and Cristina España-Bonet. 2019. Massive vs. curated word embeddings for low-resourced languages. the case of Yorùbá and Twi. *CoRR*, abs/1912.02481.

Ali Araabi and Christof Monz. 2020. Optimizing transformer for low-resource neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3429–3435, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Paul Azunre, Salomey Osei, Salomey Addo, Lawrence Asamoah Adu-Gyamfi, Stephen Moore, Bernard Adabankah, Bernard Opoku, Clara Asare-Nyarko, Samuel Nyarko, Cynthia Amoaba, Esther Dansoa Appiah, Felix Akwerh, Richard Nii Lante Lawson, Joel Budu, Emmanuel Debrah, Nana Boateng, Wisdom Ofori, Edwin Buabeng-Munkoh, Franklin Adjei, Isaac Kojo Essel Ampomah, Joseph Otoo, Reindorf Borkor, Standylove Birago Mensah, Lucien Mensah, Mark Amoako Marcel, Anokye Acheampong Amponsah, and James Ben Hayfron-Acquah. 2021. Contextual text embeddings for Twi. *CoRR*, abs/2103.15963.

Christine Basta, Marta R. Costa-jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39, Florence, Italy. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Andrew Caines. 2019. The geographic diversity of NLP conferences.

Isaac Caswell, Julia Kreutzer, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Javier Ortiz Suárez, Iroro Orife, Kelechi Ogueji, Rubungo Andre Niyongabo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Balli, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2021. Quality at a glance: An audit of web-crawled multilingual datasets. *CoRR*, abs/2103.12028.

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.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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.

Bonaventure F. P. Dossou and Mohammed Sabry. 2021. AfriVEC: Word embedding models for African languages. case study of Fon and Nobiin. *CoRR*, abs/2103.05132.

David M. Eberhard, Gary F. Simons, and Charles D. Fenning. 2019. Ethnologue: Languages of the worlds. (twenty second edition).

Ignatius Ezeani, Ikechukwu Onyenwe, and Mark Hepple. 2018. Transferred embeddings for Igbo similarity, analogy, and diacritic restoration tasks. In *Proceedings of the Third Workshop on Semantic Deep Learning*, pages 30–38, Santa Fe, New Mexico. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond English-centric multilingual machine translation. *CoRR*, abs/2010.11125.

Michael A. Hedderich, David Adelani, Dawei Zhu, Jesujoba Alabi, Udia Markus, and Dietrich Klakow. 2020. 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)*, pages 2580–2591, Online. 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.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2019. Cross-lingual ability of multilingual BERT: an empirical study. *CoRR*, abs/1912.07840.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *CoRR*, abs/2001.08361.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W. Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in Adam. *CoRR*, abs/1711.05101.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Vincent Micheli, Martin d'Hoffschmidt, and François Fleuret. 2020. On the importance of pre-training data volume for compact language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7853–7858, Online. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

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. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Kelechi Ogueji and Orevaoghene Ahia. 2019. Pidgin-UNMT: Unsupervised neural machine translation from West African Pidgin to English. *CoRR*, abs/1912.03444.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages

2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Technical report, OpenAI.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Zirui Wang, Zachary C. Lipton, and Yulia Tsvetkov. 2020. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450, Online. Association for Computational Linguistics.

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 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

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. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yian Zhang, Alex Warstadt, Haau-Sing Li, and Samuel R. Bowman. 2020. When do you need billions of words of pretraining data? *CoRR*, abs/2011.04946.