Mini But Mighty: Efficient Multilingual Pretraining with Linguistically-Informed Data Selection

Tolúlọpẹ́ Ògúnrèmí
Stanford University
tolulope@cs.stanford.edu

Dan Jurafsky
Stanford University
jurafsky@stanford.edu

Christopher D. Manning
Stanford University
manning@cs.stanford.edu

Abstract

With the prominence of large pretrained language models, low-resource languages are rarely modelled monolingually and become victims of the "curse of multilinguality" in massively multilingual models. Recently, AfriBERTa showed that training transformer models from scratch on 1GB of data from many unrelated African languages outperforms massively multilingual models on downstream NLP tasks. Here we extend this direction, focusing on the use of related languages. We propose that training on smaller amounts of data but from related languages could match the performance of models trained on large, unrelated data. We test our hypothesis on the Niger-Congo family and its Bantu and Volta-Niger sub-families, pretraining models with data solely from Niger-Congo languages and finetuning on 4 downstream tasks: NER, part-of-speech tagging, sentiment analysis and text classification. We find that models trained on genetically related languages achieve equal performance on downstream tasks in low-resource languages despite using less training data. We recommend selecting training data based on language-relatedness when pretraining language models for low-resource languages.

1 Introduction

Since the introduction of the large pretrained language models (Devlin et al., 2019), low-resource languages have not had the opportunity to be treated in the same way as high-resource languages such as English, French or Mandarin Chinese. Massively multilingual models trained using a mixture of high and low-resource languages such as mBERT (Devlin et al., 2019), XLM-RoBERTa (Conneau et al., 2020) or mT5, (Xue et al., 2021) have been proposed as a solution. Yet these do not work as well on low-resource languages as they do on high-resource languages due to the *"curse of multilinguality"* (Conneau et al., 2020), where an increase of languages in a model leads to capacity dilution, negatively affecting performance for all languages. This makes massively multilingual models sub-optimal solutions for such languages.

The quality of the training data for low-resource languages seems to differ greatly to that of high-resource languages (Kreutzer et al., 2022). The AfriBERTa models (Ogueji et al., 2021) demonstrate the considerable success of pretrained representations when trained with a 'small' (1GB), high-quality dataset focused on eleven languages of a single continent – Africa. AfriBERTa Large outperforms the larger, massively multilingual models on named-entity-recognition (NER) and text classification for various African languages. While this *continental* approach is promising, it uses a mixture of different language families that are not genetically related.

Here, we propose using language relatedness in lieu of general geographic proximity of languages to pretrain transformer models. We test this hypothesis by grouping training data by language family and then testing on four tasks: NER, Part-of-Speech Tagging (POS Tagging), Sentiment Analysis and Text Classification. New models trained range from 100 to 600 MB of training data, in contrast to 1GB of data for AfriBERTa and 2395 GB for XLM-R. We find that the smallest models trained on the most closely-related languages perform as well as models trained with up to 10 times the amount of data (AfriBERTa).

In this paper we:

- Train and release[1] pretrained models on genetically grouped African languages
- Finetune and release models for NER, POS tagging, sentiment analysis and text classification on various African languages

[1]Models are available to download at https://github.com/Tolulope/mini-but-mighty

Findings of the Association for Computational Linguistics: EACL 2023, pages 1251–1266
May 2-6, 2023

- Find that training on genetically grouped languages performs equally to larger models, despite training on much less data.

2 Related Work

Despite a long history of work on individual NLP tasks on African languages (Adedjouma Sèmiyou et al., 2012; Dibitso et al., 2019; Schlunz et al., 2016; Pauw et al., 2006; Onyenwe et al., 2014; Hunegnaw et al., 2021; Orimaye et al., 2012; Eiselen, 2016; Alabi et al., 2020), the lack of freely available and aggregated models made it difficult for languages to build off of each other.

The lack of adequate training data in low-resource languages, including African languages, led to multilingual pretraining transformer models such as mBERT (Devlin et al., 2019), XLM-RoBERTa (Conneau et al., 2020) and mT5 (Xue et al., 2021) using multilingual resources such as Wikipedia and the Common Crawl corpus.

In contrast, the "small data" approach, introduced with the release of the AfriBERTa models (Ogueji et al., 2021) advocates for pretraining models with small amounts of data solely in low-resource languages. The AfriBERTa Large model outperforms XLM-R and mBERT on text classification and NER for a few African languages. This is likely due to the lack of inclusion of a range of African language data and the use of unclean, crawled datasets in the original training data for the large models.

Our proposal to use small, high-quality data draws on the finding that small data perform competitively given the right quality of data (Kreutzer et al., 2022). Our work asks how far we can extend this small data approach by seeing whether large uncurated datasets can be outperformed or at least equalled by small, carefully selected high-quality datasets.

3 Method

3.1 Languages

In our work, we train models with a wide variety of African languages. To compare with AfriBERTa, we use the Afro-Asiatic languages (Amharic, Hausa, Somali, Tigrinya, and Afaan Oromoo) and when focussing on linguistic typology, we work on Niger-Congo Languages. The Niger-Congo family, introduced by Greenberg in 1949, is a genetic family of languages merging the Bantu and 'Semi-Bantu' families, due to the similarities found between both (Greenberg, 1949). It spans sub-Saharan African and is a genetic grouping. Figure 1 displays a heatmap of the average of the syntactic and phonological distances between the languages used extracted from WALS using lang2vec (Littell et al., 2017). We see clusters of similarity for the genetically grouped Volta-Niger and Bantu languages, and so our groups, while designed genetically, also are typologically coherent.

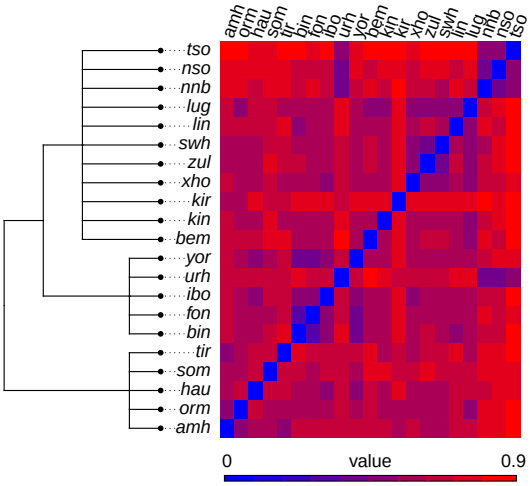


Figure 1: Heatmap displaying the average of syntactic and phonological distances queried from lang2vec between languages used to train the models along with the phylogenetic tree of the languages. Blue represents very close languages and red very distant languages. Clusters are visible for Volta-Niger languages (urh, yor, ibo, fon, and bin) and Bantu languages (nnb, nso, and tso amongst others).

The African languages used to train models in this work are summarised with their language families in Table 1.

3.2 Training Data

When training the pretrained models, we add to the AfriBERTa corpus (Ogueji et al., 2021) by collecting various data sources online. See the list of data sources in Appenedix A.1. We prioritise datasets produced solely by or in partnership with members of their communities.

3.3 Model Architecture and Training Details

We train all new models with the same architecture as AfriBERTa Large, with 6 attention heads, 768 hidden units, 3072 feedforward size, and a maximum length of 512 (Ogueji et al., 2021). Models trained from scratch are trained for 460,000

Language	ISO Code	Language Family	Branch
Afaan Oromoo	orm	Afro-Asiatic	-
Amharic	amh	Afro-Asiatic	-
Hausa	hau	Afro-Asiatic	-
Somali	som	Afro-Asiatic	-
Tigrinya	tir	Afro-Asiatic	-
Bemba	bem	Niger-Congo	Bantu
Gahuza	kir+kin	Niger-Congo	Bantu
isiXhosa	xho	Niger-Congo	Bantu
isiZulu	zul	Niger-Congo	Bantu
Kiswahili	swa	Niger-Congo	Bantu
Lingala	lin	Niger-Congo	Bantu
Luganda	lug	Niger-Congo	Bantu
Nande	nnb	Niger-Congo	Bantu
Sepedi	nso	Niger-Congo	Bantu
Setswana	ssw	Niger-Congo	Bantu
Xitsonga	tso	Niger-Congo	Bantu
Èdó	bin	Niger-Congo	Volta-Niger
Fon	fon	Niger-Congo	Volta-Niger
Igbo	ibo	Niger-Congo	Volta-Niger
Urhobo	urh	Niger-Congo	Volta-Niger
Yorùbá	yor	Niger-Congo	Volta-Niger
Nigerian Pidgin	pcm	English Creole	

Table 1: Summary of languages used for training language models with their language family, branch and ISO 639-3 code used to refer to languages in Section 4.

steps with a learning rate of $1e{-}4$. To compare pretrained to continued pretraining, we continue pretraining of the AfriBERTa model by 180,000 steps with all the data from the Niger-Congo family. We also compare newly trained models to monolingual and massively multilingual models trained with much more data: BERT Cased (Devlin et al., 2019), BERT Uncased, RoBERTa (Liu et al., 2019), and XLM-RoBERTa (Conneau et al., 2019).

To initially compare genetics with geography, we train two models with different subsets of the AfriBERTa corpus. ***AfriBERTa (Niger-Congo)*** is trained with data from the Niger-Congo languages in AfriBERTa (Gahuza, Igbo, Kiswahili and Yorùbá) and ***AfriBERTa (Afro-Asiatic)*** is trained with the Afro-Asiatic languages in AfrBERTa (Afaan Oromoo, Amharic, Hausa, Somali, and Tigrinya).

The Niger-Congo family has many branches. Due to data availability, we focus on the Volta-Niger and Bantu branches. We supplement the existing data in the AfriBERTa corpus with data in Bemba, Edo, Fon, isiXhosa, isiZulu, Kiswahili (Congolese variant), Lingala, Luganda, Nande, Sepedi, Setswana, Urhobo and Xistonga). Data from these languages totals roughly 364 MB of data. We call the model trained with all of these languages ***Niger-Congo BERTa***. We then divide the data by language family and pretrain ***BantuBERTa*** and ***VoltaBERTa***.

To test the effects of tokenisation, we train a custom tokenizer with the training data from the Niger-Congo family with the same vocabulary size as AfriBERTa, namely 70,000. The training data for the tokenizer was sampled using the method introduced in XLM (Conneau and Lample, 2019), using an $\alpha = 0.3$.

3.3.1 Size comparison models

To test whether the data selection for the Niger-Congo BERTa models results in the models' performance downstream, we train AfriBERTa models with the same amount of training data in Niger-Congo BERTa (364 MB), BantuBERTa (260 MB) and VoltaBERTa (107 MB). The resulting models are Afriberta 107, AfriBERTa 260 and AfriBERTa 364, which will be finetuned and directly compared to a model of the same size. To achieve this we proportionally reduce the amount of training data for each language in the AfriBERTa corpus to create three pretraining corpora with 107MB, 260MB and 364MB accordingly each with data from the eleven languages used to train AfriBERTa. The results are averaged across relevant languages for each sized model: Volta-Niger languages for AfriBERTa 107, Bantu languages for AfriBERTa 260 and all Niger-Congo languages for AfriBERTa 364.

The newly trained models along with AfriBERTa are summarised in Table 2.

3.4 Evaluation Data

We evaluate our models on four downstream tasks: named-entity recognition, part-of-speech tagging, sentiment analysis and text classification.

NER: For NER, we use the MasakaNER dataset (Adelani et al., 2021b), covering 10 African Languages covering Afro-Asiatic (Amharic, Hausa, Luo) and Niger-Congo languages. The Niger-Congo branches represented are Bantu (Kiyarwanda, Luganda, Kiswahili), Volta-Niger (Igbo, Yorùbá) and West Atlantic (Wolof).

POS Tagging: For POS Tagging, we use high-quality POS tagging data provided by Masakhane (which is not yet publicly available) covering Bambara, Hausa, Igbo, Kinyarwanda, Nyanja (or Chichewa), Nigerian Pidgin English, Kiswahili, isiXhosa, isiZulu and data from the DHASA-SACAIR 1st Joint Task on Part-of-Speech Tagging for African Languages covering isiNdebele, isiXhosa, isiZulu and Setswana.

Model	Languages	Training Data (MB)	Evaluation Data (MB)	Time to train (hrs)
AfriBERTa (Ogueji et al., 2021)	orm, amh, kin, kir, hau, ibo, pcm, som, swa, tir, yor	939	80	–
AfriBERTa (Niger-Congo)	kin, kir, ibo, swa and yor	279	23	57
AfriBERTa (Afro-Asiatic)	All Afro-Asiatic languages	611	57	60
AfriBERTa Continued	All Niger-Congo languages	364	41	75
Niger-Congo BERTa	All Niger-Congo languages	364	41	75
BantuBERTa	All Bantu languages	260	36	57
VoltaBERTa	All Volta-Niger languages only	107	12	57
AfriBERTa 107	orm, amh, kin, kir, hau, ibo, pcm, som, swa, tir, yor	107	12	57
AfriBERTa 260	orm, amh, kin, kir, hau, ibo, pcm, som, swa, tir, yor	260	36	57
AfriBERTa 364	orm, amh, kin, kir, hau, ibo, pcm, som, swa, tir, yor	364	41	75

Table 2: Summary of models trained and/or used in experiments. Models trained on NVIDIA TITAN RTX GPUs

Sentiment Analysis: For Sentiment Analysis, we use YOSM (Shode et al., 2022) and NaijaSenti (Muhammad et al., 2022). YOSM is a sentiment corpus of film reviews in Yorùbá. NaijaSenti is a Twitter sentiment analysis corpus covering the Nigeran languages Hausa, Igbo, Nigerian Pidgin English and Yorùbá.

Text classification: For text classification, we use a Hausa and Yorùbá news topic classification dataset (Hedderich et al., 2020) and the KINNEWS and KIRNEWS dataset (Niyongabo et al., 2020) covering Kinyarwanda and Kirundi.

4 Results

Results for our experiments are listed in Figure 2 and Tables 3 to 8. Given that datasets have data for languages in different families and branches, we select relevant models for comparison here and leave the full set of the results in the Appendix.

4.1 NER

We finetune the pretrained language models for NER using the Masakhane NER dataset. The results for the AfriBERTa model are taken from the paper (Ogueji et al., 2021). The results for our NER experiments are in Figure 2.

For Niger-Congo languages, shown in Figure 2a, ***Niger-Congo BERTa*** performs almost as well as AfriBERTa and AfriBERTa with continued pretraining. The difference in results is not statistically significant, but the slight increase may suggest that more training data results in better performance for the NER task.

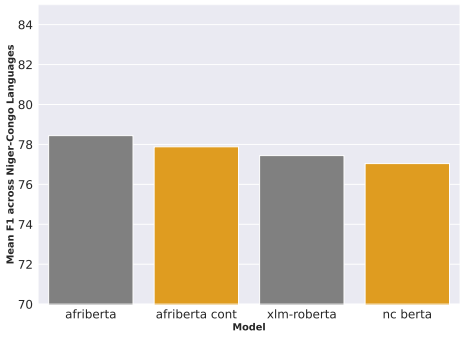
For Afro-Asiatic languages, shown in Figure 2b, the ***AfriBERTa (Afro-Asiatic)*** model performs almost as well as AfriBERTa with differences in F1 that are not statistically different (less than 0.1 F1). This suggests that training data selection based on genetic grouping results in downstream performance that is not significantly different, despite the reduction in data used. XLM-RobBERTa performs best for Luo and Nigerian Pidgin. Nigeran Pidgin is an English Creole, so we can assume the abundance of English training data in XLM-RoBERTa's training data helps performance. Luo, a language not present in the training data of any of the models has the best performance with XLM-RoBERTa. This suggests that for unseen languages and English Creoles, it may still be best to finetune massively multilingual models.

4.2 POS Tagging

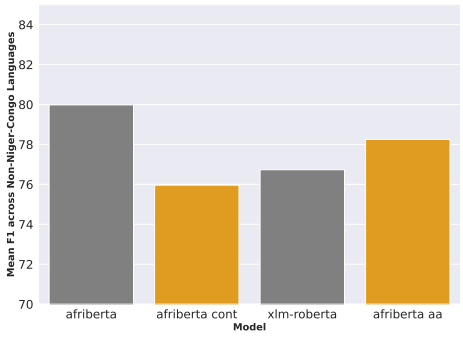
We finetune the models trained on the Part-of-Speech Tagging task, using our two datasets. With languages that have multiple datasets, we train separate models and report the mean per language.

In Table 3 we can see that ***BantuBERTa*** performs best on most Bantu languages, with an improvement on AfriBERTa of 1.8 F1 for isiZulu, 1.5 F1 for isiXhosa and 1.51 F1 for Chichewa, despite using roughly 25% of the training data of AfriBERTa. Despite the results not being significantly different, we see that training smaller models with higher quality data and a criterion of genetic relatedness leads to performance that is as good as larger models.

For Hausa, an Afro-Asiatic language, we see in Table 4 that ***AfriBERTa (Afro-Asiatic)*** does not perform significantly differently from AfriBERTa, with only a slight difference in F1 score (0.08 F1 less than AfriBERTa). This suggests that for POS Tagging, linguistically-informed data selection leads to performance that is as good as that



(a) Mean F1 of Niger-Congo languages



(b) Mean F1 of non-Niger-Congo languages

Figure 2: Plots of the mean F1 scores across languages for NER. Plot (a) shows the mean F1 for Niger-Congo languages and Plot (b) shows the mean across non-Nige-Congo languages.

lang	family	afriberta	bantu	nc berta
bam	bantu	86.65	86.70	86.8
ibo	volta-niger	80.47	78.53	80.03
kin	bantu	94.44	**94.45**	94.38
nbl	bantu	80.91	80.88	81.05
nya	bantu	81.14	**82.65**	81.67
ssw	bantu	84.50	85.44	**85.54**
swa	bantu	**92.06**	91.79	91.51
xho	bantu	84.96	**86.46**	86.40
zul	bantu	82.35	**84.16**	83.50
mean		**86.04**	85.86	85.87

Table 3: F1 scores for POS Tagging models of languages that are in the Niger-Congo family. ***BantuBERTa*** and ***Niger-Congo BERTa*** perform as well as AfriBERTa across languages.

lang	afriberta	afriberta cont	afriberta aa	xlm roberta
hau	**91.34**	89.55	91.26	89.89
pcm	87.57	86.31	86.19	**89.76**
mean	89.46	87.93	88.73	**89.83**

Table 4: F1 scores for POS Tagging models of languages that are not in the Niger-Congo family. ***Afri-BERTa (Afro-Asiatic)*** is performing almost as well as AfriBERTa for Hausa, despite being trained with much less data.

of larger models outside the Niger-Congo family. Nigerian Pidgin performs best with XLM-RoBERTa, an expected result given that Nigerian Pidgin is an English Creole.

4.3 Sentiment Analysis

The results for Yorùbá presented are the mean F1 scores from the YOSM and NaijaSenti models.

lang	afriberta	nc berta	afribera nc	volta niger
ibo	86.78	87.53	86.96	**88.48**
yor	86.09	85.92	85.93	**86.42**
mean	86.44	86.72	86.44	**87.45**

Table 5: F1 scores for Sentiment Analysis models of languages that are in the Volta-Niger family. ***VoltaBERTa*** performs as well as AfriBERTa, despite being trained with 10% of the data.

lang	afriberta	nc berta tok	xlm roberta	afriberta aa
hau	87.42	85.54	85.85	**87.43**
pcm	72.94	74.83	**79.06**	70.95
mean	80.18	80.19	**82.46**	79.19

Table 6: F1 scores for Sentiment Analysis models of languages that are not in the Volta-Niger family. ***Afri-BERTa (Afro-Asiatic)*** is performing almost as well as AfriBERTa for Hausa, despite being trained with much less data.

For Volta-Niger languages, the model trained on only 100MB of data, ***VoltaBERTa*** has the best performance for both Igbo and Yorùbá, outperforming AfriBERTa by 1.7 and 0.33 F1 despite being trained on 10% of the data. Here we see the advantages of a model being trained on a smaller, yet distinct branch of the Niger-Congo family. The results imply that a smaller linguistically-selected model is as good as a larger non-linguistically-selected model, and has the advantage of being smaller and

therefore more widely usable. It is possible that the high similarity of these languages leads to the model's increased ability learn about the languages and perform better downstream.

Hausa has the best performance with ***AfriBERTa (Afro-Asiatic)*** and Nigerian Pidgin English with XLM-Roberta. We also see that the English Creole performs best when finetuned on a model trained on English data, supporting our language-relatedness claim with a different set of languages. Training data from similar languages suffices for competitive performance downstream.

4.4 Text Classification

For text classification, we continue to see the trend that models trained on much less data do not have significantly different performance downstream. ***AfriBERTa (Afro-Asiatic)***'s performance is almost as good as AfriBERTa's for Hausa, ***BantuBERTa*** with a Niger-Congo tokenizer performs almost as well for Kinyarwanda and outperforms AfriBERTa for Kirundi for Bantu languages and ***VoltaBERTa*** does not perform significantly differently for Yorùbá. In yet another task, we demonstrate that linguistically-informed data selection trumps data quantity.

4.5 Is quality still relevant if we hold size constant?

In addition to comparing model performance with different amounts of training data, we also directly compare models trained with the same amount of data but with different sets of languages with varying levels of genetic similarity below.

Table 8 summarises the training data experiments with the mean F1 score for each model across languages for each task. AfriBERTa 107 is compared to the ***VoltaBERTa*** model as they both use 107 MB of training data, AfriBERTa 260 is compared to the ***BantuBERTa*** as both models use 260 MB of data and AfriBERTa 364 is compared to ***Niger-Congo BERTa*** as they both use 364 MB of training data. We train the ***Niger-Congo BERTa*** models with and without a custom tokenizer. The results from models trained with a custom tokezier have an asterisk. We see that when size is held constant the models trained with high-quality data from closely-related languages perform at least as well as models train with data from a wider range of languages. These results highlight the importance of data selection when resources are limited and support our claim that pretraining with genetically-related languages doesn't result in significantly different performance downstream.

Overall, we see that across tasks and languages, models trained with data from genetically related languages alone work as well as models trained with up to 10 times the amount of data.

5 Model Visualisation

5.1 Model Visualisation

To visualise the models, we extract sentence embeddings by concatenating the weights of the last four layers of the model for 1,000 sentences in each language's evaluation set. We use 1000 sentences for each language to ensure an even distribution across languages. For dimensionality reduction, we use Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) and visualise each sentence in two dimensions. We present UMAP plots as they are not as sensitive to parameters as t-SNE.

Visualisations of models grouped by language family (specific branches when part of the Niger-Congo family) are below. All visualisations show evidence of language-specific and family-specific clustering in the models.

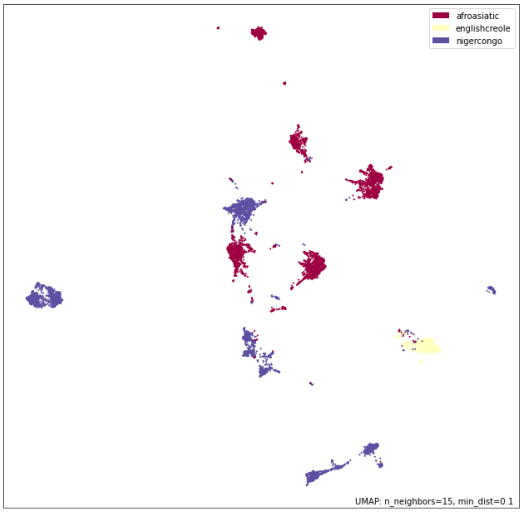


Figure 3: AfriBERTa visualised with languages in the training, coloured by language family (Afro-asiatic in pink, English Creole in yellow and Niger-Congo languages in purple). There appear to be language specific clusters.

When reduced by UMAP, AfriBERTa does not seem to cluster languages by family. Nigerian Pidgin English, is situated away from most of other languages, apart from Yorùbá (bottom right). This

lang	family	afriberta	afriberta cont	afriberta aa	bantu tok	nc berta tok	volta niger
hau	afro-asiatic	**90.13**	88.18	89.84	84.1	84.22	71.77
kin	bantu	73.87	**74.41**	70.45	73.69	73.46	68.26
kir	bantu	82.37	84.18	81.38	**84.72**	83.59	80.91
yor	volta-niger	79.88	**80.63**	70.88	70.52	78.98	79.70
mean		81.56	**81.85**	78.14	78.24	80.06	76.77

Table 7: F1 scores for Text Classification models

	afriberta 107	volta niger	afriberta 260	bantu	afriberta 364	nc berta
NER	79.61	**82.46**	78.10	**79.45**	76.66	**77.04**
POS Tagging	79.32	**80.40**	85.51	**86.56**	84.78	**85.67***
Sentiment Analysis	85.30	**87.45**				
Text Classification	76.68	**78.15**	77.50	**79.21***	78.11	**78.68***
mean per model	80.23	**82.12**	80.23	**81.74**	79.85	**80.46**

Table 8: Table showing the mean F1 across languages in each sub-family compared to an AfriBERTa model trained on the same amount of data for each task. Results with an asterisk (*) are from models trained with a custom tokenizer.

is could be due to borrowing of Yorùbá words into Nigerian Pidgin English.

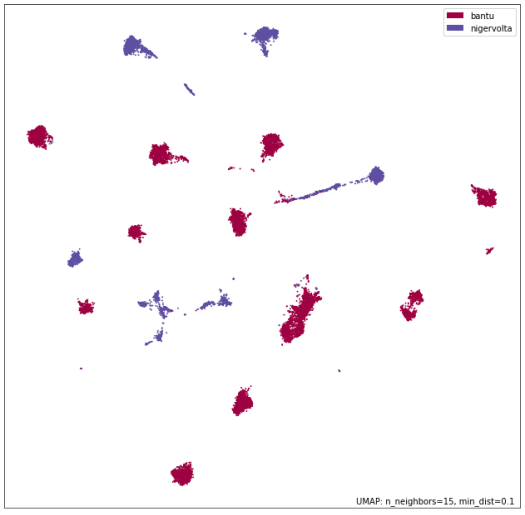


Figure 4: Niger-Congo BERTa visualised with languages in the training, coloured by language family (Bantu languages in pink and Volta-Niger languages in purple). We see language-specific clusters, but no branch-specific separation of the language clusters.

The ***Niger-Congo BERTa*** model does not seem to cluster languages by sub-family. This may be because all the languages are in the same larger family already.

The ***VoltaBERTa*** model completely splits Bantu and Volta-Niger Languages, possibly helped by the absence of Bantu languages in the training data. This could be due to the scripts of Igbo and Yorùbá

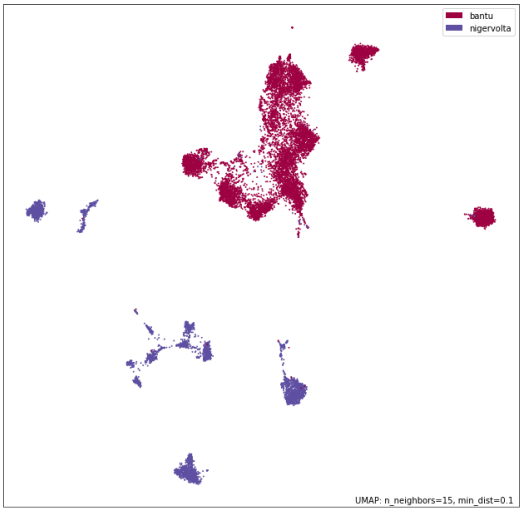


Figure 5: ***VoltaBERTa*** visualised with languages in the training, coloured by language family (Bantu languages in pink and Volta-Niger languages in purple). Here, Bantu languages are clearly separated from Volta-Niger languages.

.

(use of diacritics) and Fon (use of different characters), leading the model to internally distinguish between languages part of the Volta-Niger family and those that are not.

Overall, we see that the more closely related the languages used to train the pretrained model, the more distinct the representations of different language families or branches in the UMAP visualisations. This is most likely due to the other languages not being present in the training data, but

the results for POS Tagging and Sentiment Analysis show that this focus on closely related languages leads to improvements in performance with much less data.

6 Discussion

In this work we see that when pretraining multilingual models with closely related languages, the resulting finetuned models work just as well as models finetuned on a wider variety of languages. Sentence embeddings show that the more closely related the languages in the training data, the better the model's ability to differentiate language families.

We do not see one model consistently outperforming others. However, we do see multilingual models of closely related languages work for those languages downstream and generalise better to unseen languages within the family. ***BantuBERTa*** works very well for POS Tagging of Bantu languages and ***VoltaBERTa*** for sentiment analysis of Volta-Niger languages. Continued pretraining of AfriBERTa with closely related languages gives the best text classification result on average. This "small data" combined with language similarity approach demonstrates that it is possible to maintain performance with fewer resources, possibly at the expense of using different models for different downstream tasks.

7 Conclusion

In this paper, we have pretrained several multilingual transformer models exclusively with low-resource languages. We have shown that the grouping of closely-related languages in training data can match or improve performance across several downstream tasks despite the reduction in training data used. We have also demonstrated that for very low-resource languages, we can exploit language similarity to improve performance of NLP tasks on these languages with models trained on similar languages only.

8 Limitations

In this work we did not have an exact overlap of downstream tasks to training data and therefore could not exactly match pretrained models to general task performance. We did not have Bantu language data for Sentiment Analysis, preventing us from making conclusions on this task with BantuBERTa. We also note that we only have data from two branches of the Niger-Congo family. Data from a wider variety of branches would have helped us make more general conclusions.

We did not compare any of our models to finetuned large language models, nor did we fine–tune our pretrained models before finetuning them for the downstream tasks. It is possible that language-adaptive finetuning of Niger-Congo languages on these models trained exclusively on Niger-Congo languages may lead to even better performance. Given the lack of resources in these languages, one would have to determine guidelines on which data would be used for pretraining or finetuning in this case.

9 Acknowledgements

We would like to thank the anonymous reviewers, Alex Tamkin, Kaitlyn Zhou and Mirac Suzgun for their comments.

This work was supported by Award IIS-2128145 from the NSF and the Stanford School of Engineering Fellowship.

References

A Adedjouma Sèmiyou, John OR Aoga, and Mamoud A Igue. 2012. Part-of-speech tagging of Yoruba standard, language of Niger-Congo family. *Research Journal of Computer and Information Technology Sciences*, 1:2–5.

David Adelani, Dana Ruiter, Jesujoba Alabi, Damilola Adebonojo, Adesina Ayeni, Mofe Adeyemi, Ayodele Esther Awokoya, and Cristina España-Bonet. 2021a. The effect of domain and diacritics in Yoruba–English neural machine translation. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 61–75, Virtual. Association for Machine Translation in the Americas.

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 Ge-

breyohannes, 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. 2021b. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Jesujoba Alabi, Kwabena Amponsah-Kaakyire, David Adelani, and Cristina España-Bonet. 2020. Massive vs. curated embeddings for low-resourced languages: the case of Yorùbá and Twi. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2754–2762, Marseille, France. European Language Resources Association.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale.

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.

MA Dibitso, PA Owolawi, and SO Ojo. 2019. Part of speech tagging for Setswana African language. In *2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, pages 1–6. IEEE.

Roald Eiselen. 2016. Government domain named entity recognition for South African languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3344–3348, Portorož, Slovenia. European Language Resources Association (ELRA).

Chris Chinenye Emezue and Femi Pancrace Bonaventure Dossou. 2020. FFR v1.1: Fon-French neural machine translation. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, pages 83–87, Seattle, USA. Association for Computational Linguistics.

Ignatius Ezeani, Paul Rayson, Ikechukwu Onyenwe, Chinedu Uchechukwu, and Mark Hepple. 2020. Igbo-english machine translation: An evaluation benchmark.

Joseph H. Greenberg. 1949. Studies in African linguistic classification: I. the Niger-Congo family. *Southwestern Journal of Anthropology*, 5(2):79–100.

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.

Ashebir Hunegnaw et al. 2021. Sentiment analysis model for Afaan Oromoo short message service text: A machine learning approach. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(13):332–342.

Julia Kreutzer, Isaac Caswell, 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 Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, 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 Ballı, 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. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. 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. *arXiv preprint arXiv:1907.11692*.

Rooweither Mabuya, Jade Abbott, and Vukosi Marivate. 2021. Umsuka english - isizulu parallel corpus.

Vukosi Marivate and Tshephisho Sefara. 2020. South african news data.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Cindy McKellar. 2018. Autshumato Setswana Monolingual Corpora.

Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Ibrahim Said Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Anuoluwapo Aremu, Saheed Abdul, and Pavel Brazdil. 2022. Naijasenti: A Nigerian twitter sentiment corpus for multilingual sentiment analysis. *arXiv preprint arXiv:2201.08277*.

Jonathan Mukiibi, Babirye Claire, and Nakatumba-Nabende Joyce. 2021. The Makerere MT Corpus: English to Luganda parallel corpus.

Rubungo Andre Niyongabo, Qu Hong, Julia Kreutzer, and Li Huang. 2020. KINNEWS and KIRNEWS: Benchmarking cross-lingual text classification for Kinyarwanda and Kirundi. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5507–5521, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. 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*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ikechukwu Onyenwe, Chinedu Uchechukwu, and Mark Hepple. 2014. Part-of-speech tagset and corpus development for Igbo, an African language. In *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*, pages 93–98, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Sylvester Olubolu Orimaye, Saadat M Alhashmi, and Siew Eu-gene. 2012. Sentiment analysis amidst ambiguities in YouTube comments on Yoruba language (Nollywood) movies. In *Proceedings of the 21st International Conference on World Wide Web*, pages 583–584.

Guy De Pauw, Gilles-Maurice de Schryver, and Peter W Wagacha. 2006. Data-driven part-of-speech tagging of Kiswahili. In *International Conference on Text, Speech and Dialogue*, pages 197–204. Springer.

Wikus Pienaar, Wildrich Fourie, and Cindy McKellar. 2018. Autshumato English-Xitsonga Manually Translated Parallel Corpora.

Georg I Schlunz, Nkosikhona Dlamini, and Rynhardt P Kruger. 2016. Part-of-speech tagging and chunking in text-to-speech synthesis for South African languages. In *Interspeech 2016*. Curran Associates, Inc.

Shivachi Casper Shikali and Mokhosi Refuoe. 2019. Language modeling data for swahili.

Iyanuoluwa Shode, David Ifeoluwa Adelani, and Anna Feldman. 2022. YOSM: A new Yorùbá sentiment corpus for movie reviews. In *3rd Workshop on African Natural Language Processing*.

Claytone Sikasote and Antonios Anastasopoulos. 2022. Bembaspeech: A speech recognition corpus for the bemba language. In *Proceedings of the Language Resources and Evaluation Conference*, pages 7277–7283, Marseille, France. European Language Resources Association.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. 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*, pages 483–498, Online. Association for Computational Linguistics.

Alp Öktem, Muhannad Albayk Jaam, Eric DeLuca, and Grace Tang. 2020. Gamayun - language technology for humanitarian response. In *2020 IEEE Global Humanitarian Technology Conference (GHTC)*, pages 1–4.

A Appendix

A.1 List of data sources

A.6 Training data comparisons

Language	Data Sources
Afaan Oromoo	AfriBERTa Corpus
Amharic	AfriBERTa Corpus
Hausa	AfriBERTa Corpus
Somali	AfriBERTa Corpus
Tigrinya	AfriBERTa Corpus
Bemba	Text from Bemba Speech Corpus (Sikasote and Anastasopoulos, 2022)
Gahuza	AfriBERTa Corpus
isiXhosa	Xhosa Navy Parallel Corpus (Tiedemann, 2012)
isiZulu	Umsuka English - isiZulu Parallel Corpus (Mabuya et al., 2021)
Kiswahili	AfriBERTa Corpus, Language modeling data for Swahili (Shikali and Refuoe, 2019) and Gamayun (Öktem et al., 2020) Congolese Kiswahili Medium kit
Lingala	Gamayun (Öktem et al., 2020) Lingala Kit
Luganda	Makerere MT Corpus (Mukiibi et al., 2021)
Nande	Gamayun (Öktem et al., 2020) Nande kit
Sepedi	South African News Data (Marivate and Sefara, 2020)
Setswana	Autshumato Setswana Monolingual Corpora (McKellar, 2018) and South African News Data (Marivate and Sefara, 2020)
Xitsonga	Autshumato English-Xitsonga Manually Translated Parallel Corpora (Pienaar et al., 2018)
Ẹ̀dó	JW300 (Agić and Vulić, 2019)
Fon	FFR Translate Corpus (Emezue and Dossou, 2020)
Igbo	AfriBERTa Corpus and Igbo Monolingual Dataset (Ezeani et al., 2020)
Urhobo	JW300 (Agić and Vulić, 2019)
Yorùbá	AfriBERTa Corpus and MENYO-20k dataset (Adelani et al., 2021a)
Nigerian Pidgin	AfriBERTa Corpus

Table 9: List of sources for language data used to train the models in Table 2.

A.2 Full NER Results

lang	afri berta nc	xlm roberta	bantu tok	volta niger tok	afri berta aa	bert cased	nc berta	nc berta tok	volta niger	bantu	afri berta cont
amh	37.9 ± 7.11	55.85± 2.45	0.0 ± 0.0	0.0 ± 0.0	**72.73**± 5.64	0.0 ± 0.0	40.09± 5.68	0.0 ± 0.0	7.95 ± 69.67	39.68± 3.1	63.11± 9.69
hau	85.0 ± 3.01	89.35± 3.0	84.26± 1.28	82.34± 3.9	**90.11**± 2.49	85.89± 3.09	84.43± 2.89	84.83± 1.25	82.31± 3.94	83.72± 2.26	87.64± 1.73
ibo	87.16± 1.86	83.96± 2.16	75.99± 2.11	86.65± 2.01	83.19± 1.44	83.13± 2.45	86.97± 3.88	86.03± 3.25	86.59± 2.4	77.5 ± 3.99	**87.44**± 2.61
kin	71.78± 4.26	72.36± 3.56	72.27± 3.99	62.71± 3.9	65.34± 3.11	71.35± 3.28	71.77± 2.8	71.17± 4.67	63.01± 2.65	72.43± 5.88	**72.47**± 5.77
lug	78.42± 2.7	80.0 ± 4.58	77.85± 1.41	70.48± 3.49	75.17± 2.75	77.82± 4.46	79.3 ± 3.2	78.21± 3.32	70.46± 3.39	78.28± 6.65	**78.97**± 4.79
luo	68.96± 3.09	74.73± 5.19	70.1 ± 5.87	58.63± 5.57	68.62± 5.93	**73.05**± 5.66	69.86± 2.2	70.06± 8.19	59.29± 9.08	67.93± 4.38	69.71± 4.81
pcm	81.18± 1.81	**86.97**± 3.12	76.05± 3.23	75.91± 5.04	81.54± 1.77	86.8 ± 5.5	80.92± 4.05	79.09± 5.46	76.17± 5.44	76.26± 4.57	83.38± 5.99
swa	87.3 ± 2.1	87.16± 2.0	87.62± 2.4	77.28± 4.18	81.49± 3.19	83.73± 2.53	86.83± 2.81	85.94± 2.27	77.33± 4.12	87.64± 1.51	**87.87**± 1.51
wol	58.37± 5.33	**64.87**± 3.95	59.15± 4.5	51.81± 9.09	59.16± 10.79	62.77± 8.6	59.54± 10.09	57.65± 10.21	52.66± 10.73	59.84± 7.7	61.43± 3.2
yor	79.04± 5.42	76.28± 6.12	68.75± 4.91	77.81± 3.17	69.79± 6.45	73.2 ± 4.29	77.85± 3.98	78.21± 6.41	78.32± 3.97	68.57± 7.66	**79.07**± 4.3

Table 10: Full set of NER Tagging Results. Models are finetuned five times with the mean and 95% confidence interval displayed.

A.3 Full Text Classification Results

	nc berta	volta niger	afriberta	bantu tok	nc berta tok	afriberta aa	afriberta cont	bantu
hau	81.08 ± 2.24	73.85 ± 7.79	**90.13** ± 2.75	84.10 ± 2.0	84.22 ± 4.85	89.84 ± 1.21	88.18 ± 2.59	78.26 ± 2.31
kin	73.2 ± 1.29	67.56 ± 3.26	73.87 ± 2.42	73.69 ± 2.26	73.46 ± 2.5	70.45 ± 1.94	74.41 ± 1.77	**74.07** ± 2.1
kir	81.28 ± 5.04	80.52 ± 1.36	82.37 ± 9.38	**84.72** ± 3.26	83.59 ± 6.31	81.38 ± 2.34	84.18 ± 2.33	82.47 ± 6.16
yor	79.69 ± 6.17	79.70 ± 3.53	79.88 ± 5.41	70.52 ± 4.56	78.98 ± 4.43	70.88 ± 8.5	**80.63** ± 3.08	69.15 ± 4.23

Table 11: Full set of the Text Classification results. Models are finetuned five times with the mean and 95% confidence interval displayed.

A.4 Full POS Tagging Results

	afri berta nc	xlm roberta	afri berta	bantu tok	volta niger tok	afri berta aa	bert cased	nc berta	nc berta tok	volta niger	bantu	afri berta cont
bam	86.66± 2.03	**88.23**± 0.4	86.97± 1.63	87.1± 2.72	86.71± 0.85	86.6± 0.95	87.79± 1.63	86.8± 0.67	87.01± 1.45	86.97± 1.45	86.7± 1.8	86.62± 1.49
hau	87.77± 1.28	90.44± 1.61	91.13± 1.0	88.39± 2.07	87.54± 1.02	**91.26**± 1.37	89.12± 2.6	87.78± 2.08	88.71± 1.38	87.23± 0.9	87.89± 0.92	89.74± 1.99
ibo	79.7± 1.71	79.99± 2.3	80.26± 3.33	77.68± 2.02	79.75± 5.21	77.73± 3.07	79.19± 1.64	80.03± 2.47	80.29± 2.55	79.88± 1.87	78.53± 2.67	**80.80**± 2.36
kin	93.91± 1.17	93.15± 1.07	94.28± 0.7	94.41± 0.51	83.8± 2.26	89.3± 2.76	93.36± 1.29	94.38± 1.18	93.97± 0.2	85.73± 2.55	**94.45**± 0.68	93.96± 1.04
nbl	80.38± 1.52	**81.83**± 0.48	80.74± 0.48	80.67± 0.58	79.34± 0.99	79.97± 1.94	81.65± 1.38	81.05± 0.98	80.74± 0.57	79.92± 1.42	80.88± 0.36	80.53± 0.96
nya	80.52± 1.65	82.03± 1.83	80.98± 1.15	81.33± 1.41	77.51± 2.27	79.71± 2.2	80.91± 3.65	81.67± 2.68	82.04± 2.78	78.39± 2.86	**82.65**± 2.92	80.96± 0.99
pcm	85.86± 0.99	**89.76**± 1.5	87.64± 1.55	85.43± 1.55	84.44± 1.45	86.19± 1.06	89.68± 1.53	85.87± 1.11	85.95± 1.47	84.55± 1.35	85.56± 1.47	86.42± 0.58
ssw	84.14± 2.4	84.89± 1.21	85.0± 1.54	85.09± 1.73	82.64± 1.06	84.25± 0.56	84.29± 1.82	**85.54**± 0.75	85.14± 1.71	83.25± 2.29	85.44± 0.74	**85.54**± 1.64
swa	92.03± 1.36	91.73± 0.99	91.59± 1.12	91.74± 1.01	84.25± 1.07	87.08± 1.55	89.77± 1.91	91.51± 0.95	91.71± 1.34	84.71± 1.31	**91.79**± 0.86	91.77± 0.81
xho	92.7± 1.49	94.52± 1.07	93.52± 1.75	94.84± 0.51	90.38± 1.24	92.85± 0.96	93.39± 0.44	94.84± 0.5	94.57± 1.01	91.79± 0.58	**95.05**± 0.39	94.5± 0.64
xho1	75.77± 2.76	77.5± 2.28	76.62± 1.89	77.82± 1.1	67.03± 3.65	74.98± 2.71	74.83± 2.91	77.97± 1.49	**78.08**± 3.4	72.92± 1.95	77.76± 2.15	78.03± 3.34
zul	84.7± 1.12	85.46± 0.18	85.21± 0.59	86.28± 0.87	83.49± 1.34	84.68± 1.51	84.65± 1.35	85.5± 1.37	85.8± 0.81	84.29± 1.9	**85.95** ± 1.42	85.26± 0.79
zul1	79.73± 1.75	82.2± 2.86	79.35± 2.05	81.78± 1.35	73.1± 3.79	77.57± 3.32	80.96± 2.34	81.49± 2.62	81.79± 1.54	76.61± 2.91	**82.36**± 1.52	82.25± 1.91

Table 12: Full set of POS Tagging Results. Models are finetuned five times with the mean and 95% confidence interval displayed.

A.5 Full Sentiment Analysis Results

lang	afri berta nc	afri berta	bert un-cased	bantu tok	xlm rob erta	volta niger tok	bert cased	mbert	nc berta	nc berta tok	volta niger	bantu
hau	84.15± 1.75	**87.42**± 1.13	81.74± 2.41	85.98± 2.44	85.85± 1.64	85.38± 1.4	84.02± 3.24	83.25± 3.63	84.35± 3.55	85.54± 1.49	82.47± 2.96	83.1± 2.31
ibo	86.96± 3.23	86.78± 1.46	80.44± 4.32	84.07± 2.22	84.62± 13.16	87.11± 1.54	85.03± 4.74	84.99± 3.92	87.53± 2.31	86.58± 3.08	**88.48**± 2.38	83.81± 2.85
pcm	70.73± 3.71	72.94± 5.13	75.02± 9.45	77.47± 8.6	**79.06**± 1.38	72.21± 1.3	73.73± 13.96	71.95± 15.06	66.59± 5.06	74.83± 4.62	63.55± 10.43	71.0± 11.9
yor	84.49± 2.02	85.18± 2.55	79.01± 3.71	82.03± 2.72	55.29± 0.0	**86.38**± 2.14	82.98± 2.57	80.96± 19.38	85.64± 0.82	85.11± 1.44	85.77± 1.18	82.48± 0.9
yosm	**87.36**± 5.48	87.0± 4.57	72.83± 4.25	80.29± 4.52	82.83± 2.35	85.79± 5.38	82.43± 2.23	83.59± 5.45	86.19± 5.11	85.99± 4.72	87.07± 4.22	76.98± 2.54

Table 13: Full set of Sentiment Analysis Results. Yorùbá data from NaijaSenti (*yor*) and Yorùbá data from YOSM (*yosm*) were finetuned separately. Models are finetuned five times with the mean and 95% confidence interval displayed.

	afriberta 107	volta niger	volta niger tok
ibo	84.45	86.59	**86.65**
yor	74.76	**78.32**	77.81
mean	79.61	**82.46**	82.23

Table 14: F1 scores for models of the same size finetuned for NER on Volta-Niger languages. ***VoltaBERTa*** performs best overall

	afriberta 260	bantu	bantu tok
kin	70.29	**72.43**	72.27
lug	77.27	**78.28**	77.85
swa	86.75	**87.64**	87.62
mean	78.10	**79.45**	79.25

Table 15: F1 scores for models of the same size finetuned for NER on Bantu languages. The ***BantuBERTa*** model outperforms the AfriBERTa model of the same size on Bantu languages by 1.35 F1 on average for NER.

	afriberta 364	nc berta	nc berta tok
ibo	86.45	**86.97**	86.03
kin	**72.43**	71.77	71.17
lug	76.09	**79.3**	78.21
swa	**87.37**	86.83	85.94
wol	**60.38**	59.54	57.65
yor	77.22	77.85	**78.21**
mean	76.66	**77.04**	76.20

Table 16: F1 scores for models of the same size finetuned for NER on Niger-Congo languages. ***Niger-Congo BERTa*** performs best on average.

	afriberta 107	volta niger
yor	76.68	**78.15**

Table 17: F1 scores for models of the same size finetuned for Text Classification on Yorùbá, a Volta-Niger languages with ***VoltaBERTa*** outperforming the AfriBERTa model trained on the same amount of data.

	afriberta 260	bantu	bantu tok
kin	73.17	**74.07**	73.69
kir	81.83	82.47	**84.72**
mean	77.5	78.27	**79.21**

Table 18: F1 scores for models of the same size finetuned for Text Classification on Bantu languages. The ***BantuBERTa*** model, both with and without a custom tokenizer outperforms the AfriBERTa model of the same size on Bantu languages.

	afriberta 364	nc berta	nc berta tok
kin	72.96	73.20	**73.46**
kir	82.29	81.28	**83.59**
yor	79.08	**79.69**	78.98
mean	78.11	78.06	**78.68**

Table 19: F1 scores for models of the same size finetuned for Text Classification on Niger-Congo languages. ***Niger-Congo BERTa*** with and without a custom tokenizer perform better than the AfriBERTa model of the same size.

	afriberta 107	volta niger	volta niger tok
ibo	79.32	**80.40**	79.75

Table 20: F1 scores for models of the same size finetuned for POS Tagging on Volta-Niger languages with ***VoltaBERTa*** outperforming the AfriBERTa model trained on the same amount of data.

	afriberta 260	bantu	bantu tok
bam	**87.26**	86.7	87.1
kin	93.64	**94.45**	94.41
nbl	80.25	**80.88**	80.67
nya	80.83	**82.65**	81.33
ssw	84.45	**85.44**	85.09
swa	91.59	**91.79**	91.74
xho	84.42	**86.41**	86.33
zul	81.64	**84.16**	84.03
mean	85.51	**86.56**	86.34

Table 21: F1 scores for models of the same size finetuned for POS Tagging on Bantu languages with ***BantuBERTa*** almost always outperforms the AfriBERTa model trained on the same amount of data.

	afriberta 364	nc berta	nc berta tok
bam	86.84	86.8	**87.01**
ibo	80.28	80.03	**80.29**
kin	93.8	**94.38**	93.97
nbl	79.89	**81.05**	80.74
nya	81.00	81.67	**82.04**
ssw	83.74	**85.54**	85.14
swa	91.74	91.51	**91.71**
xho	84.03	**86.41**	86.33
zul	81.74	83.50	**83.80**
mean	84.78	85.65	**85.67**

Table 22: F1 scores for models of the same size finetuned for POS Tagging on Niger-Congo languages. ***Niger-Congo BERTa*** with and without a custom tokenizer perform better than the AfriBERTa model of the same size.

	afriberta 107	volta niger	volta niger tok
ibo	85.85	**88.48**	87.11
yor	84.74	**86.42**	86.09
mean	85.30	**87.45**	86.60

Table 23: F1 scores for models of the same size fine-tuned for Sentiment Analysis on Volta-Niger languages with ***VoltaBERTa*** consistently outperforming the Afri-BERTa model trained on the same amount of data.