YoNER: A new Yorùbá Multi-domain Named Entity Recognition Dataset

Peace Busola Falola[1], **Jesujoba O. Alabi**[2], **Solomon O. Akinola**[1], **Folashade T. Ogunajo**[3], **Emmanuel Oluwadunsin Alabi**[1], **David Ifeoluwa Adelani**[5]
[1]University of Ibadan, Nigeria, [2]Saarland University, Germany, [3]Atiba University, Nigeria,
[5]Mila - Quebec AI Institute, McGill University, and Canada CIFAR AI Chair
peacefalola@gmail.com, david.adelani@mcgill.ca

Abstract

Named Entity Recognition (NER) is a foundational NLP task, yet research in Yorùbá has been constrained by limited and domain-specific resources. Existing resources, such as MasakhaNER (a manually annotated news-domain corpus) and WikiAnn (automatically created from Wikipedia), are valuable but restricted in domain coverage. To address this gap, we present YoNER, a new multidomain Yorùbá NER dataset that extends entity coverage beyond news and Wikipedia. The dataset comprises about 5,000 sentences and 100,000 tokens collected from five domains including Bible, Blogs, Movies, Radio broadcast and Wikipedia, and annotated with three entity types: Person (PER), Organization (ORG) and Location (LOC), following CoNLL-style guidelines. Annotation was conducted manually by three native Yorùbá speakers, with an inter-annotator agreement of over 0.70, ensuring high quality and consistency. We benchmark several transformer encoder models using cross-domain experiments with MasakhaNER 2.0, and we also assess the effect of few-shot in-domain data using YoNER and cross-lingual setups with English datasets. Our results show that African-centric models outperform general multilingual models for Yorùbá, but cross-domain performance drops substantially, particularly for blogs and movie domains. Furthermore, we observed that closely related formal domains, such as news and Wikipedia, transfer more effectively. In addition, we introduce a new Yorùbá-specific language model (OyoBERT) that outperforms multilingual models in in-domain evaluation. We publicly release the YoNER dataset and pretrained OyoBERT models to support future research on Yorùbá natural language processing.

Keywords: Named Entity Recognition, Low-resource Language, African NLP, Multidomain NER, Yorùbá

1. Introduction

Named Entity Recognition (NER) is a foundational NLP task that identifies and classifies named entities (e.g. personal name, organizations, locations) in text with several applications in information extraction, question answering, and speech recognition (Tjong Kim Sang and De Meulder, 2003; Yamada et al., 2020; Caubrière et al., 2020). To enable broad applicability, NER models are developed for diverse domains (Weischedel et al., 2011), and techniques to facilitate faster adaptation across multiple domains remain an active area of research (Yang and Katiyar, 2020; Das et al., 2022b; Ashok and Lipton, 2023; Xue et al., 2024; Huang et al., 2025). While there are several *multi-domain* datasets available for high-resource languages such as English, they are often lacking for low-resource languages, where existing datasets are typically limited to Wikipedia or news domains (Palen-Michel et al., 2024).

For low-resource languages such as Yorùbá, spoken by over 50 million people in Nigeria and neighboring African countries, NER is crucial for information extraction and also play an important function of recognizing African names and cultural concepts which are often unrecognized by current AI systems (Olatunji et al., 2023). Recently, there have been some efforts to create human-annotated NER datasets for Yorùbá but they are mostly in the "NEWS" domain (Alabi et al., 2020; Adelani et al., 2021, 2022). Outside NEWS, the other available data is WikiANN based on Wikipedia, however, the annotations are automatically created and only 100 test samples and few entities.

In this paper, we develop the first multi-domain Yorùbá NER dataset known as **YoNER**. YONER covers five distinct domains including: *Bible*, *Blogs*, *Movies*, *Radio* broadcast, and *Wikipedia*. Each domain has between 800 and 1400 sentences collected from various sources by a Yorùbá native speaker. We then recruited three native speakers for the annotation and quality control. Leveraging YONER, we attempt to answer the following research questions: (1) How well does NEWS NER data transfer to diverse domains? (2) How well can small in-domain data e.g. 200 sentences improve domain transfer? (3) Does monolingual BERT models enhance multi-domain transfer compared to multilingual BERT models in low-resource setting? To answer the third question, we pre-trained a new Yorùbá BERT model (OYO-BERT) and compare the performance to African-centric BERT models such as AfroXLMR (Alabi et al., 2022). Finally, we extend this evaluation to English language to see if the same result holds true, by comparing English

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7137–7148
11–16 May 2026.

RoBERTa-Large with XLM-Roberta-Large.

Our evaluation shows adapting NER model to other domains is quite challenging especially for the Blog and Movie domain, often with fewer entities. The best adaptation was from News to Wikipedia. Moreover, adapting from Yorùbá News domain achieved better performance than English CoNLL NER data with almost twice of its training data, this shows that both source language and closeness of source domains to target are equally important. We obtained a better adaptation by combining the small in-domain data from all sources and the News domain to achieve the most impressive results. While the newly developed OyoBERT seem to achieve better performance than AfroXLMR, it does not seem to achieve better cross-domain transfer results.[1][2]

2. Related Work

NER for low-resource languages has recently gained growing attention as the development of multilingual models such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) has enabled effective transfer learning across languages. Traditional NER systems relied heavily on large annotated datasets, which are scarce for most low-resource languages. Multilingual pre-trained language models have helped bridge this gap by transferring knowledge from high-resource languages to creating competitive models in zero and few-shot setting (Zhao et al., 2021; Adelani et al., 2021; Schmidt et al., 2022). This progress has opened new opportunities for languages that were previously under-served in NLP research.

In the African context with several low-resource languages, significant progress has been achieved through community-driven initiatives aimed at building foundational datasets and benchmarks. The first effort was by Alabi et al. (2020) that developed Yorùbá NER data based on Global Voices news articles, with 26K tokens. The MasakhaNER project (MasakhaNER 1.0 and MasakhaNER 2.0) created larger resources (Adelani et al., 2021, 2022) with 83K and 244K tokens (between 3K and 10K sentences) respectively. The MasakhaNER datasets represents high-quality, manually annotated NER datasets for over 21 African languages from the news domain, including Yorùbá. Other NER datasets available are WikiAnn (Pan et al., 2017) and NaijaNER (Oyewusi et al., 2021). The former is automatically annotated from Wikipedia by projecting annotations from English, however the size is small (100 sentences test set) while the latter is also based on news domain but they expanded the entity types from four (PER, ORG, LOC and DATE) to 18 entity types covered by OntoNotes (Weischedel et al., 2011). Existing datasets do not cover diverse domains, and we address this gap in this paper.

Cross-domain NER is an active research area, as models trained on one genre typically perform poorly on others due to shifts in vocabulary and entity distributions (Jia et al., 2019; Liu et al., 2021). While there have been efforts to develop new methods and datasets to address this issue (Jia et al., 2019), most of these have focused on high-resource languages such as English, leaving low-resource languages relatively underexplored and underrepresented in cross-domain NER research.

Our work contributes to this line by providing multi-domain human annotated data for Yorùbá called **YoNER**. Given **YoNER**, we conducted several experiments to study the cross-domain transfer across the domains using several transformer based language models.

3. Corpus Curation for YoNER

3.1. Data Collection

We compiled Yorùbá texts from five domains, aiming for varied style and content. We focused on popular domains covered in previous works such as OntoNotes and WikiAnn: *Bible*, *Blog*, *Radio*, *Wikipedia*, *Movie*. For the two domains, movies and radio, with which we were less familiar, we collaborated with a Yoruba newscaster to collect sentences. More details about the data collection process are provided below.

Bible: We collected verses from the Yorùbá Bible (Bible.com corpora). Based on the authors familiarity with the Bible, we chose books and chapters that contained more entities such as `Genesis 10`, `Genesis 14`, `Exodus 6`, `Numbers 1`, `Numbers 3`, `Numbers 26`, `Joshua 12`, `Joshua 18`, `I Samuel 8:1`, `I Samuel 22`, `II Samuel 5`, `I Kings 9`, `Nehemiah 3`, `Ezra 2:1`, `Psalms 87`, `Jeremiah 25`, `Joel 3`, `Matthew 2`, `Matthew 23`, `Luke 2`, `Acts 6`, `Acts 15` and `Acts 27`.

Blogs: We curated posts and comments from Yorùbá blogs and forums (e.g. Nairaland, social media pages, and Yoruba news blogs). This domain includes informal language, slang, and possible code-mixing.

Movies: We collected transcripts of dialogues from Yorùbá movies (collected from YouTube subtitled content). Movie dialogue often contains conversational Yorùbá, interjections, and may omit diacritics. Sentences were gathered from four movies for the movies domain. The movies are `Oleku`, `Apoti Eri`, `Jagunjagun` and `Owo Eje`, all ob-

[1]We release YoNER publicly on Hugging Face.

[2]We release OyoBERT models publicly: YoBERT-base, OyoBERT-base, OyoBERT-large.

Domain	Yorùbá Sentence
Bible	Àwọn ọmọkùnrin Simeoni , Jemueli , Jamini , Ohadi , Jakini , Sohari , Saulu , tí ìyá rẹ jẹ́ ọmọbìnrin arábìnrin Kenaani .
Blog	Ipò obinrin kò pin si idi àdìrò àti inú ilé yòkù gẹ́gẹ́ bi Olórí Òṣèlú Nigeria , Muhammadu Buhari ti sọ.
Movie	Wòó Àṣàké , jẹ́ ká nasẹ̀ lọ, ká lọ gba'tẹ́gùn díẹ̀.
Radio	Ìràwọ̀ agbábọ́ọ̀lù tìnú ẹgbẹ́ ìkọ̀ Liverpool Salah nló tí ńlérí wí pé ẹgbẹ́ agbábọ́ọ̀lù náà yóò gba ìfẹ̀ẹ̀ yẹ ní sáà tó ńbọ̀.
Wikipedia	Òun ló gba àmì ẹ̀yẹ PEN Pinter Prize ní ọdún 2018 Ilẹ́ẹ̀kọ́ sẹ́kọ́ndìrì Yunifásítì ti Naijiria , Nsukka ni Adichie ti parí ẹ̀kọ́ sẹ́kọ́ndìrì rẹ̀ níbi tí ó ti ọ̀pọ̀lọpọ̀ àmì ẹ̀yẹ ajemọ́ akadá.

Table 1: Examples of named entities across sentences from different Yorùbá domains in YoNER. (PER , ORG , LOC).

tained from YouTube and Netflix. All movies except Apoti Eri was code-mixed. A Yorùbá newscaster manually transcribed all the movies.

Radio: We collected transcripts and recordings from radio stations broadcasting in Yorùbá. (e.g. Tiwa-n-Tiwa, Fresh FM). These include news and talk shows, blending formal and colloquial speech.

Wikipedia (Wiki): We make use of Yorùbá Wikipedia articles obtained from HuggingFace.[3] This domain is encyclopedic and covers a broad range of topics but uses a formal written style.

Table 1 presents an example sentence for each domain covered in YoNER, along with the corresponding annotations. The examples include annotated personal names and locations from different domains, such as a list of biblical names in the Bible domain and the mention of a Nigerian political figure in a blog.

Data split After cleaning and sentence segmentation of the collected data, the sentences were divided into splits with final counts of 200 for training, 100 for development, and the remaining sentences for testing (ranging from 500 to 1,128 test sentences per domain). In total, the corpus contains 5,148 sentences (100,795 tokens) across domains. Table 2 shows the statistics of the dataset.

3.2. Annotation Process

We recruited three native Yorùbá speakers (who are also authors on this paper) to annotate the collected sentences of various domains. Annotation was done via the Human Signal annotation platform (formerly Label Studio). Annotators were trained on the NER annotation guidelines used by MasakhaNER (Adelani et al., 2021) to label each token for three entity types: PER (person), LOC (location), and ORG (organization). We used the IOB2 scheme for spans: each token is tagged with B-, I-, or O (outside any entity).

Each sentence was annotated independently by all three annotators.

Domain	Source	Train/dev/test	Tokens
Blogs	Nairaland, Yoruba blog, Akonilekede Yoruba, Facebook	200/100/553	20,123
Bible	Bible.com	200/100/713	23,896
Movies	YouTube	200/100/1128	14,312
Radio	Tiwa-n-Tiwa and Fresh FM Radio Stations	200/100/752	21,404
Wikipedia	Wikipedia	200/100/502	21,090
Total		**5,148**	**100,795**
MasakhaNER 2.0	Existing	6877/983/1964	244,144

Table 2: YoNER Dataset statistics by domain.

Inter-Annotator Agreement After labeling, we computed inter-annotator agreement using Fleiss' Kappa at two levels: token level (ignoring B/I differences, just whether each token is labeled as an entity or not) and entity level (treating each contiguous span as an entity annotation). The results in Table 3 show substantial agreement at the token level (k=0.71–0.88) but much lower at the entity level (k=0.30–0.61). This gap indicates that annotators often agreed on which words are entities but had more disagreement on exact span boundaries (a common issue in NER). Disagreements were resolved by majority vote (or discussion when

[3] https://huggingface.co/datasets/wikimedia/wikipedia

Domain	Fleiss' κ (token)	Fleiss' κ (entity)
Blogs	0.7064	0.3957
Bible	0.8775	0.5938
Movies	0.7911	0.3038
Radio	0.8313	0.6072
Wikipedia	0.8190	0.5758

Table 3: Inter-annotator agreement (Fleiss' κ) across Yoruba NER domains.

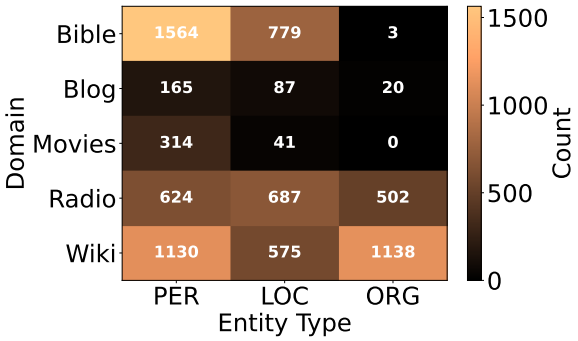


Figure 1: Frequency of Entity Types By Domain in YoNER dataset.

all three annotators differed) to produce the final gold-standard labels.

The final annotated dataset is therefore high-quality Yoruba NER data across five domains. Figure 1 shows the entity counts across each domain. Across all domains, PER entities are the most frequent, while ORG entities are rare, appearing mainly in data from Radio and Wikipedia.

4. Experimental Setup

Given the created YoNER dataset, in this section we outline our experimental setup designed to answer the research questions posed earlier.

4.1. Cross-Domain Transfer of News NER

To address the first research question, "*How well does News NER data transfer to diverse domains?*," we trained NER models on data from the News domain and evaluated them on other domains. Specifically, we fine-tuned several multilingual transformer models on the Yorùbá portion of the MasakhaNER 2.0 (Adelani et al., 2022) dataset, which primarily consists of news text. We selected MasakhaNER 2.0 because of its moderately large size (over 6,000 training sentences) and its rich collection of African geographical and cultural entities, making it a suitable source domain for evaluating cross-domain transfer in Yorùbá. The resulting models were then evaluated on the test splits of each domain in the YoNER dataset.

Models: The selected multilingual PLMs include AfriBERTa (Ogueji et al., 2021), AfriBERTa-V2 (Oladipo, 2024), mBERT (Devlin et al., 2019), XLM-R (base and large; (Conneau et al., 2020)), and AfroXLMR (base, large, and large-76L; (Alabi et al., 2022; Adelani et al., 2024)).

4.2. Impact of Small In-Domain Data on NER Transfer

To address the second research question, "*How well can small in-domain data (e.g., 200 sentences) improve domain transfer?*," we fine-tuned the best-performing base model from the previous experiment on the training split of each domain in YoNER. The resulting models were then evaluated both in-domain and out-of-domain to assess how a small amount of domain-specific data affects NER performance across diverse text domains. To provide a cross-lingual baseline for comparison, we also included English NER datasets covering domains similar to those in YoNER. Specifically, we used CoNLL and WikiAnn (Pan et al., 2017) for the News and Wikipedia domains, respectively, while from the OntoNotes (Weischedel et al., 2011) dataset, we selected broadcast conversations and web blogs to represent the Radio and Blogs domains. In this case, all available English data for each domain were used for training, and the Bible domain was included only in the evaluation phase, as no corresponding English training data were available. Furthermore, all the English datasets were processed to have only PER, LOC and ORG as entity types, all other entity types were changed to O except for OntoNotes where we changed GPE to LOC.

4.3. Monolingual vs. Multilingual Models for Multi-Domain Transfer

Previous research has shown that multilingual transformer models often underperform compared to their monolingual counterparts across different languages (Rönnqvist et al., 2019; Pyysalo et al., 2021). We investigate this observation in our third research question. To address the question, "*Do monolingual BERT models enhance multi-domain transfer compared to multilingual BERT models?*," we compare Named Entity Recognition (NER) models trained in English using RoBERTa (a monolingual model) with XLM-R (a multilingual model). For Yorùbá, since no well-established monolingual models are available but multilingual models exist, we develop a BERT model for Yorùbá, as described in Section 4.4. We fine-tuned the BERT model on each individual domain and evaluated it on the

Model	Size	Topic Class.	News Topic Class.	Twitter Sentiment	Movie Sentiment	NER	POS	Avg.
Baselines								
AfriBERTa (Ogueji et al., 2021)	126M	70.6	90.3	72.9	90.9	87.7	94.5	84.5
AfroXLMR-large (Alabi et al., 2022)	550M	74.8	94.0	74.1	84.3	**89.3**	**95.0**	85.3
AfroXLMR-large-76L (Adelani et al., 2024)	550M	79.9	**94.7**	**75.1**	88.0	88.4	95.2	86.8
YoBERT-base (ours)	110M	69.2	92.1	71.1	90.0	85.0	94.0	83.6
OyoBERT-base (ours)	110M	80.5	94.0	73.0	90.6	86.7	94.4	86.5
OyoBERT-large (ours)	337M	**82.5**	93.8	74.9	**92.6**	87.4	94.6	**87.6**

Table 4: Evaluation of OyoBERT models and other multilingual baselines on four sequence-level and two token-level classification tasks for Yorùbá.

corresponding test set, comparing the results to fine-tuning on a combination of all five domains simultaneously. For the YoNER experiments, we compared BERT with AfroXLMR-large-76L under this setup. We replicated the same procedure for English as well. To replicate a similar low-resource setting as in YoNER, where we had only 200 training examples per domain, our initial experiments on the English datasets showed that approximately 1,000 examples and relatively long sentences were needed to achieve reasonable in-domain F1 scores. Therefore, for the English experiment, we randomly selected 1,000 sentences from CoNLL, WikiAnn, and OntoNotes (as described in the previous section) containing more than five tokens, and compared RoBERTa-large with XLMR-large.

4.4. Oyo-BERT pre-training

To address the third question, we pre-trained a BERT model for Yorùbá. We trained a BERT model with token dropping objective (Hou et al., 2022)—where unimportant tokens are dropped in the intermediate layer but later picked up by the last layer so that the model still produces full-length sequences. We make use of an open-source implementation.[4] The BERT model was trained on a TPU v3-8 Google cloud compute in less than 24 hours.

The pre-training data is based on the Yorùbá subset of mC4—pre-training corpus for mT5 (Xue et al., 2021) (153MB), MT560 (Gowda et al., 2021) (59MB), Yorùbá portion of the Wikipedia (11MB), and a number of curated news sources such as BBC Yorùbá (15MB), Alaroye (10MB), Awinkoko (7MB), and Asejere (2MB). The total size of the monolingual data was around 347MB.

To further increase training data, we leverage synthetic data obtained by machine translating additional contents for English, similar to how AfroXLMR-76L was created (Adelani et al., 2024)–where languages less than 10MB leverage MT generated data from NLLB-200 (600M parameters) (NLLB-Team et al., 2022).[5]. Here, we translated over 1GB of English texts from WURA (Oladipo et al., 2023)—An high-quality African corpus containing English to Yorùbá with NLLB (600M) (NLLB Team et al., 2024). This increased the entire pre-training corpus to over 1.93GB.

The resulting models are trained for two model sizes: The **OyoBERT-base** has the same configuration as BERT-base while the **OyoBERT-Large** has same configuration as BERT-large. The BERT trained without MT data is referred to as **YoBERT**.

4.5. Model Hyper-parameters

For all NER fine-tuning experiments, we adapted an open-source codebase.[6] We used a maximum sequence length of 256, a batch size of 32, a learning rate of 5e-5, and trained for 50 epochs. For evaluation, we report the micro-averaged F1 score.

5. Result and Discussion

In this section, we present the results of our experiments, beginning with the evaluation of the newly created PLMs for Yorùbá. We then proceed to address the research questions outlined earlier.

5.1. How well does OyoBERT perform on wide range of tasks?

Here, we compared the performance of the *newly* trained Yorùbá BERT models (YoBERT, OyoBERT-base and OyoBERT-large) to well established multilingual baselines such as AfriBERTa (Ogueji et al., 2021), AfroXLMR (Alabi et al., 2022) and AfroXLMR-76L (Adelani et al., 2024) covering 11, 20 and 76 languages respectively including Yorùbá.

We evaluated on six representative NLP evaluation data including SIB-200 (Adelani et al., 2024)

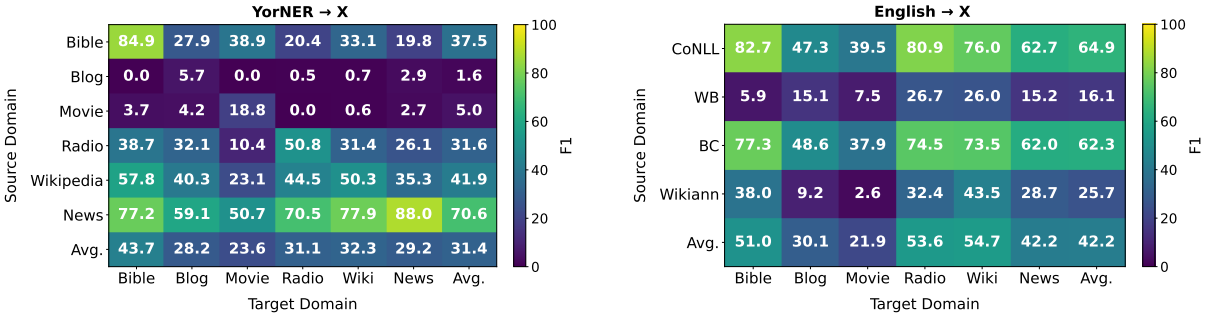
[4] https://github.com/stefan-it/model-garden-lms

[5] Important to note that Yorùbá did not use synthetic data in AfroXLMR since it has more than 300MB

[6] https://github.com/masakhane-io/masakhane-ner/tree/main/MasakhaNER2.0

Model	Bible	Blog	Movies	Radio	Wiki	News	Avg.
MasakhaNER 2.0							
OyoBERT-large	$69.80_{1.10}$	$54.30_{1.23}$	$42.17_{1.53}$	$72.50_{3.03}$	$74.14_{1.31}$	$87.78_{0.49}$	66.78
AfriBERTa	$58.85_{1.72}$	$52.77_{1.18}$	$44.47_{1.76}$	$\mathbf{75.26}_{0.72}$	$68.96_{0.95}$	$87.42_{0.58}$	64.62
AfriBERTa-V2	$71.01_{1.22}$	$50.38_{2.10}$	$47.12_{1.84}$	$76.47_{2.11}$	$75.01_{0.45}$	$87.22_{0.29}$	67.87
mBERT	$61.94_{1.54}$	$43.38_{3.71}$	$25.72_{4.34}$	$76.10_{1.22}$	$70.10_{0.63}$	$84.30_{0.35}$	60.26
XLM-R-base	$63.20_{1.21}$	$40.22_{1.54}$	$42.40_{3.47}$	$68.54_{1.52}$	$66.28_{1.18}$	$84.19_{0.51}$	60.81
XLM-R-large	$68.80_{2.24}$	$42.24_{2.26}$	$42.19_{5.03}$	$68.38_{2.19}$	$67.62_{1.86}$	$84.12_{0.88}$	62.23
AfroXLMR-base	$73.90_{1.58}$	$50.71_{1.56}$	$39.63_{2.63}$	$72.63_{1.64}$	$73.88_{1.03}$	$86.59_{0.32}$	66.22
AfroXLMR-large	$74.48_{0.72}$	$58.50_{1.17}$	$49.19_{5.83}$	$71.36_{1.23}$	$77.35_{0.80}$	$\mathbf{88.77}_{0.42}$	69.94
AfroXLMR-large-76L	$\mathbf{77.22}_{2.39}$	$\mathbf{59.07}_{3.33}$	$\mathbf{50.69}_{2.61}$	$70.49_{2.06}$	$\mathbf{77.88}_{1.91}$	$88.04_{0.73}$	**70.57**

Table 5: F1 score (%) for cross-domain transfer to the YoNER dataset from the News domain. Scores are averaged over five runs. All models are fine-tuned on the training split of MasakhaNER 2.0.



(a) YoNER → X

(b) English → X

Figure 2: Cross-domain evaluation of NER model trained on each YoNER domain and on domain-specific English NER datasets. English-trained models are evaluated on YoNER domains to assess cross-lingual transfer. Average over 5 runs.

(a topic classification dataset to categorize sentences), MasakhaNEWS (Adelani et al., 2023) (a news topic classification to categorize news articles), AfriSenti (Muhammad et al., 2023) (a Twitter sentiment classification dataset), NollySenti (Shode et al., 2023) (a movie sentiment classification dataset), MasakhaNER (Adelani et al., 2021) (for NER), and MasakhaPOS (Dione et al., 2023) (for part-of-speech tagging).

Table 4 shows the average evaluation of OyoBERT on six tasks. First observation is that it outperform the YoBERT with about $+3$ points, this shows the benefit of more data, specifically synthetic data generated by MT models. Unsurprisingly, the larger OyoBERT-large achieved better results than the OyoBERT-base. When compared to the strong baselines, OyoBERT-base with only 110M parameters performed better than AfriBERTa with 126M parameters. Similarly, OyoBERT-large with only 337M parameters achieved better overall performance to AfroXLMR-76L models with 550M parameters. In general, OyoBERT has comparable or better performance on the sequence classification tasks, however, they struggle with token classification task such as NER, where AfroXLMR-large and AfroXLMR-76 achieved $+1.9$ and $+1.0$ better performance.

Overall, OyoBERT-large is the first strong monolingual encoder model for Yorùbá with comparable performance to other Africa-centric models but perform slightly worse on NER.

5.2. How well does News NER data transfer to diverse domains?

Table 5 presents the results obtained for all models when fine-tuned on MasakhaNER 2.0 and evaluated on each domain in YoNER, including the MasakhaNER 2.0 test set, denoted as "`News`". Overall, while all models achieved at least 84% F1 when the `News` model was evaluated on its corresponding test set, we observed a drop in F1 across all other YoNER domains. In particular, AfroXLMR-large-76L obtained the highest average F1, reaching 88.04% on News. The next best domains were Bible and Wikipedia, scoring 77.22% and 77.88%, respectively, which represents a drop of nearly 11%. Blog and Movies showed the lowest cross-domain performance from the News model, with F1 scores of 59.07% and 50.69%, respectively. This is also consistent with the fact that Blog and Movies are the domains with the fewest annotated entities (as shown in Figure 1).

Furthermore, a look at the average performance

of all the models shows that, although many are multilingual except OyoBERT, the African-centric models consistently perform better for cross-domain transfer, while mBERT and XLM-R (base and large) are the least performing models. We attribute the low performance of XLM-R to the fact that Yorùbá was not part of it's pre-training languages, while for mBERT, the Yorùbá portion is based on Wikipedia which is less than 30k articles.

Overall, African-centric models outperform general multilingual models for Yorùbá, but cross-domain performance drops substantially, especially for Blog and Movies.

5.3. How well can small in-domain data improve domain transfer?

Figure 2a shows the cross-domain performance of NER models trained on each domain in YoNER, based on AfroXLMR-large-76L, the best model from the previous experiment. The figure also includes results from Table 4 and the evaluation of YoNER models on MasakhaNER 2.0. The diagonal of the figure represents in-domain performance. As observed, Bible has the highest in-domain performance with 84.9%, followed by Radio and Wikipedia with 50.8% and 50.3%, respectively, while Blog and Movies exhibit the lowest in-domain performance.

Regarding cross-domain transfer, aside from News (MasakhaNER 2.0), Wikipedia is the best source domain, followed by Bible and Radio. The best target domain is Bible, followed by Wikipedia and Radio. However, all YoNER domains fail to transfer effectively to the News domain, even though the News model transfers significantly better to these domains. We attribute this to the large training size of MasakhaNER 2.0, which gives the News model a strong both in-domain and cross-domain performance.

Overall, closely related domains such as News and Wikipedia with formally written texts transfers better to each other.

5.4. How well does Cross-lingual transfer compares to cross-domain transfer?

Figure 2b shows the result from the cross-lingual experiment. For cross-lingual transfer from English datasets, CoNLL, WB, and BC did not perform best when transferred to their corresponding Yorùbá domains (News, Blog, and Radio), with only WikiAnn transferring best to Wikipedia. Considering cross-lingual cross-domain transfer, on average, CoNLL and BC were the best source domains, while WB got the lowest F1 score. As target domains, Wikipedia and Radio achieved the highest performance, whereas Blog and Movies were the worst-performing targets. Although Bible is not the best target domain on average, it achieved the highest F1 scores specifically from CoNLL and BC as source domains. In several cases, cross-lingual transfer from English datasets outperforms training on just 200 in-domain examples from YoNER, both for in-domain and cross-domain evaluations. There large data for cross-lingual may be more effective than small available in-domain data.

However, when we compare fine-tuning on large Yorùbá news NER data (MasakhaNER 2.0) with 6.8K sentences and English news NER data (CoNLL03) with 14.9K, we found the former setting to be more effective in adapting to other domains, achieving 70.6 average performance (Figure 2a) while English to other domains achieved 64.9 (+5.7 points improvement). This finding shows that transfer within the same language is easier than from another language, despite the difference in domains.

Overall, cross-domain transfer within the same language is more effective than cross-lingual transfer especially with large training source data.

5.5. Does monolingual BERT models enhance multi-domain transfer compared to multilingual BERT models in low-resource setting?

Table 6 shows the in-domain and multi-domain performance of the YoNER datasets, comparing OyoBERT-large and AfroXLMR-large-76L. We observed that, in in-domain evaluation, the language-specific model on average outperforms the multi-

Model	Bible	Blog	Movies	Radio	Wiki	Avg.
In-domain						
OyoBERT-large	81.47	**7.78**	**34.06**	**65.76**	43.09	**46.43**
AfroXLMR-large	**84.85**	5.67	18.78	50.82	**50.26**	42.08
Multi-domain						
OyoBERT-large	85.00	71.53	**56.10**	89.60	76.37	75.72
AfroXLMR-large	**89.63**	**76.09**	53.72	**90.84**	**79.64**	**77.98**

Table 6: F1 score (%) for In-domain evaluation of each of the YoNER domains. Average is over 5 runs.

Model	CoNLL	WB	BC	Wikiann	Avg.
In-domain (all data)					
RoBERTa-large	93.14	**64.37**	88.16	82.48	82.04
XLM-R-large	**93.70**	63.12	**91.37**	**82.74**	**82.73**
In-domain (1000 examples)					
RoBERTa-large	**74.33**	**1.79**	**40.98**	**54.14**	**42.81**
XLM-R-large	53.59	0.00	18.23	40.32	28.04
Multi-domain (1000 examples)					
RoBERTa-large	82.36	48.77	69.58	64.55	66.32
XLM-R-large	**83.52**	**49.44**	**74.50**	**62.91**	**67.59**

Table 7: F1 score (%) for In-domain evaluation of each of the English domains. Average is over 5 runs.

lingual model. Looking at the individual in-domain results, OyoBERT outperforms AfroXLMR in culturally specific domains such as Movies and Radio, but is less competitive on Bible and Wiki, which are not culturally specific domains. For the multi-domain evaluation, AfroXLMR appears to benefit from exposure to more data and, on average, outperforms OyoBERT; however, it still underperforms on the Movies domain. This underscores that language-specific models can be competitive in certain contexts, particularly within culturally specific domains.

Replicating this same experiment in English and English datasets, comparing RoBERTa-large and XLM-R-large (as shown in Table 7), we observed findings similar to those from YoNER. The results show that, for in-domain evaluation, the language-specific model RoBERTa-large outperforms its multilingual counterpart. However, under the multi-domain setup, XLM-R-large gains an advantage, likely due to its exposure to a broader range of training data.

Overall, in low-data regime, monolingual BERT is more beneficial for cross-domain transfer, however with more training data, multilingual BERT-based models win.

6. Which entity types do models tend to perform poorly on?

To answer this question, we examined the entity type F1 scores for the multidomain models described in Section 5.5 and presented in Table 8. Our results show that for both OyoBERT and AfroXLMR-large, the ORG entity type consistently has the lowest F1 score. Furthermore, for the Bible and Movies domains, there are no ORG entities in the dataset, as illustrated in Figure 1. We also found that for the Blog and Radio domains, the models perform better at identifying LOC entities than other types, whereas for both Bible and Wikipedia, the models tend to perform best on PER entities.

Overall, these results highlight that model performance varies significantly across domains and entity types, with a consistent weakness in recognizing ORG entities.

7. Conclusion

In this paper, we introduce YoNER, a multi-domain human annotated NER dataset Yorùbá that extends entity coverage beyond news and Wikipedia. The dataset comprises about 5,000 sentences and 100,000 tokens collected from five domains including Bible, Blogs, Movies, Radio broadcast, and Wikipedia, and annotating them for person, organization, and location entities. We benchmark several transformer encoder models using cross-domain experiments with MasakhaNER 2.0, and we also assess the effect of few-shot in-domain data using YoNER and cross-lingual setups with English datasets. Our results show that African-centric models outperform general multilingual models for Yorùbá, but cross-domain performance drops substantially, particularly for blogs and movie domains. Furthermore, we observed that closely related formal domains, such as news and Wikipedia, transfer more effectively.

Domain	PER	LOC	ORG
OyoBERT-large			
Bible	0.89	0.80	0.00
Blog	0.70	0.80	0.42
Movies	0.61	0.41	0.00
Radio	0.88	0.95	0.80
Wiki	0.87	0.84	0.53
AfroXLMR-large			
Bible	0.91	0.83	0.00
Blog	0.66	0.82	0.57
Movies	0.50	0.37	0.00
Radio	0.89	0.96	0.89
Wiki	0.83	0.84	0.66

Table 8: F1 score (%) for multi-domain training on YoNER. Average is over 5 runs.

Given the small size of the training split in YoNER, one aspect we did not explore but which remains important is few-shot NER. However, many recent approaches to few-shot learning require a language-specific model (Ding et al., 2021; Das et al., 2022a). In this paper, we created a Yoruba-specific model in OyoBERT, and we hope that future work will build upon it and YoNER to advance few-shot NER research for Yorùbá and other low-resource African languages.

8. Limitation

In this paper, we introduce YoNER, a moderately large multi-domain NER dataset covering five domains. Three of these domains contain little to no ORG entities due to their nature, and none include DATE entities. We hope that future work can extend YoNER to additional domains and a broader range of entity types. We did not evaluate large language models, which are now ubiquitous and general-purpose, and we hope that future work will explore their performance on YoNER.

Lastly, a limitation of this work is that the pretrained Yorùbá encoder models were trained on relatively small corpora, including translated texts. While they achieve strong performance on the evaluated downstream tasks, we do not perform an extensive analysis of their pretraining quality or repre-

sentations, which would be necessary to fully understand their capabilities. Future work should therefore examine the properties captured by these representations, evaluate potential biases introduced by translated data, and explore scaling pretraining with larger and more diverse Yorùbá corpora. Such investigations would provide a clearer picture of the models' generalization ability and their suitability for broader downstream applications.

9. Acknowledgment

David Adelani acknowledges the funding of Natural Sciences and Engineering Research Council (NSERC) of Canada, IVADO and the Canada First Research Excellence Fund.

10. Bibliographical References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Selasi Osei, et al. 2021. Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

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

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure FP Dossou, Akintunde Oladipo, Doreen Nixdorf, et al. 2023. Masakhanews: News topic classification for african languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 144–159.

David Ifeoluwa Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba O. Alabi, Shamsuddeen H. Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Elvis Mboning, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo L. Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Adeyemi, Gilles Q. Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu Ngoli, and Dietrich Klakow. 2022. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jesujoba Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to african languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349.

Jesujoba Alabi, Kwabena Amponsah-Kaakyire, David Ifeoluwa Adelani, and Cristina España-Bonet. 2020. Massive vs. curated embeddings for low-resourced languages: The case of yorùbá and twi. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC)*, pages 2754–2762, Marseille, France.

Dhananjay Ashok and Zachary C Lipton. 2023. Promptner: Prompting for named entity recognition. *arXiv preprint arXiv:2305.15444*.

Antoine Caubrière, Sophie Rosset, Yannick Estève, Antoine Laurent, and Emmanuel Morin. 2020. Where are we in named entity recognition from speech? In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4514–4520, Marseille, France. European Language Resources Association.

Sarkar Snigdha Sarathi Das, Arzoo Katiyar, Rebecca Passonneau, and Rui Zhang. 2022a. CONTaiNER: Few-shot named entity recognition via contrastive learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6338–6353, Dublin, Ireland. Association for Computational Linguistics.

Sarkar Snigdha Sarathi Das, Arzoo Katiyar, Rebecca J Passonneau, and Rui Zhang. 2022b. Container: Few-shot named entity recognition via contrastive learning. In *Proceedings of the*

60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6338–6353.

Ning Ding, Guangwei Xu, Yulin Chen, Xiaobin Wang, Xu Han, Pengjun Xie, Haitao Zheng, and Zhiyuan Liu. 2021. Few-NERD: A few-shot named entity recognition dataset. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3198–3213, Online. Association for Computational Linguistics.

Cheikh M Bamba Dione, David Ifeoluwa Adelani, Peter Nabende, Jesujoba Alabi, Thapelo Sindane, Happy Buzaaba, Shamsuddeen Hassan Muhammad, Chris Chinenye Emezue, Perez Ogayo, Anuoluwapo Aremu, et al. 2023. Masakhapos: Part-of-speech tagging for typologically diverse african languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10883–10900.

Thamme Gowda, Zhao Zhang, Chris Mattmann, and Jonathan May. 2021. Many-to-English machine translation tools, data, and pretrained models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 306–316, Online. Association for Computational Linguistics.

Le Hou, Richard Yuanzhe Pang, Tianyi Zhou, Yuexin Wu, Xinying Song, Xiaodan Song, and Denny Zhou. 2022. Token dropping for efficient BERT pretraining. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3774–3784, Dublin, Ireland. Association for Computational Linguistics.

Li Huang, Haowen Liu, Qiang Gao, Jiajing Yu, Guisong Liu, and Xueqin Chen. 2025. Adversity-aware few-shot named entity recognition via augmentation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24132–24140.

Chen Jia, Xiaobo Liang, and Yue Zhang. 2019. Cross-domain NER using cross-domain language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2464–2474, Florence, Italy. Association for Computational Linguistics.

Zhengbao Liu, Yichong Xu, Tianyi Yu, Wenlong Dai, Ziwei Ji, Samuel Cahyawijaya, Pascale Fung, et al. 2021. Crossner: Evaluating cross-domain named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13452–13460, Virtual.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif Mohammad, Sebastian Ruder, et al. 2023. Afrisenti: A twitter sentiment analysis benchmark for african languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13968–13981.

NLLB-Team, Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

NLLB Team et al. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841.

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.

Akintunde Oladipo, Mofetoluwa Adeyemi, Orevaoghene Ahia, Abraham Toluwalase Owodunni, Odunayo Ogundepo, David Ifeoluwa Adelani, and Jimmy Lin. 2023. Better quality pre-training data and t5 models for African languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 158–168, Singapore. Association for Computational Linguistics.

Tobi Olatunji, Tejumade Afonja, Bonaventure FP Dossou, Atnafu Lambebo Tonja, Chris Chinenye Emezue, Amina Mardiyyah Rufai, and Sahib Singh. 2023. Afrinames: Most asr models. In *Proc. Interspeech 2023*, pages 5077–5081.

Chester Palen-Michel, Maxwell Pickering, Maya Kruse, Jonne Sälevä, and Constantine Lignos. 2024. Openner 1.0: Standardized open-access named entity recognition datasets in 50+ languages. *arXiv preprint arXiv:2412.09587*.

Sampo Pyysalo, Jenna Kanerva, Antti Virtanen, and Filip Ginter. 2021. WikiBERT models: Deep

transfer learning for many languages. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 1–10, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Samuel Rönnqvist, Jenna Kanerva, Tapio Salakoski, and Filip Ginter. 2019. Is multilingual BERT fluent in language generation? In *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, pages 29–36, Turku, Finland. Linköping University Electronic Press.

Fabian David Schmidt, Ivan Vulić, and Goran Glavaš. 2022. Don't stop fine-tuning: On training regimes for few-shot cross-lingual transfer with multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10725–10742, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Iyanuoluwa Shode, David Ifeoluwa Adelani, Jing Peng, and Anna Feldman. 2023. Nollysenti: Leveraging transfer learning and machine translation for nigerian movie sentiment classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 986–998.

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

Ralph Weischedel, Eduard Hovy, Mitchell Marcus, Martha Palmer, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue. 2011. Ontonotes: A large training corpus for enhanced processing. *Handbook of Natural Language Processing and Machine Translation. Springer*, 3(3):3–4.

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.

Xiaojun Xue, Chunxia Zhang, Tianxiang Xu, and Zhendong Niu. 2024. Robust few-shot named entity recognition with boundary discrimination and correlation purification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19341–19349.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

Yi Yang and Arzoo Katiyar. 2020. Simple and effective few-shot named entity recognition with structured nearest neighbor learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6365–6375.

Mengjie Zhao, Yi Zhu, Ehsan Shareghi, Ivan Vulić, Roi Reichart, Anna Korhonen, and Hinrich Schütze. 2021. A closer look at few-shot crosslingual transfer: The choice of shots matters. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5751–5767, Online. Association for Computational Linguistics.

11. Language Resource References

Adelani, David Ifeoluwa and Abbott, Jade and Neubig, Graham and D'souza, Daniel and Kreutzer, Julia and Lignos, Constantine and Palen-Michel, Chester and Buzaaba, Happy and Rijhwani, Shruti and Ruder, Sebastian and Mayhew, Stephen and Azime, Israel Abebe and Muhammad, Shamsuddeen H. and Emezue, Chris Chinenye and Nakatumba-Nabende, Joyce and Ogayo, Perez and Anuoluwapo, Aremu and Gitau, Catherine and Mbaye, Derguene and Alabi, Jesujoba and Yimam, Seid Muhie and Gwadabe, Tajuddeen Rabiu and Ezeani, Ignatius and Niyongabo, Rubungo Andre and Mukiibi, Jonathan and Otiende, Verrah and Orife, Iroro and David, Davis and Ngom, Samba and Adewumi, Tosin and Rayson, Paul and Adeyemi, Mofetoluwa and Muriuki, Gerald and Anebi, Emmanuel and Chukwuneke, Chiamaka and Odu, Nkiruka and Wairagala, Eric Peter and Oyerinde, Samuel and Siro, Clemencia and Bateesa, Tobius Saul and Oloyede, Temilola and Wambui, Yvonne and Akinode, Victor and Nabagereka, Deborah and Katusiime, Maurice and Awokoya, Ayodele and MBOUP, Mouhamadane and Gebreyohannes, Dibora and Tilaye, Henok and Nwaike, Kelechi and Wolde, Degaga and Faye, Abdoulaye and Sibanda, Blessing and Ahia, Orevaoghene and Dossou, Bonaventure F. P. and

Ogueji, Kelechi and DIOP, Thierno Ibrahima and Diallo, Abdoulaye and Akinfaderin, Adewale and Marengereke, Tendai and Osei, Salomey. 2021. *MasakhaNER: Named Entity Recognition for African Languages*. MIT Press.

Adelani, David Ifeoluwa and Liu, Hannah and Shen, Xiaoyu and Vassilyev, Nikita and Alabi, Jesujoba O. and Mao, Yanke and Gao, Haonan and Lee, En-Shiun Annie. 2024. *SIB-200: A Simple, Inclusive, and Big Evaluation Dataset for Topic Classification in 200+ Languages and Dialects*. Association for Computational Linguistics.

Adelani, David Ifeoluwa and Neubig, Graham and Ruder, Sebastian and Rijhwani, Shruti and Beukman, Michael and Palen-Michel, Chester and Lignos, Constantine and Alabi, Jesujoba O. and Muhammad, Shamsuddeen H. and Nabende, Peter and Dione, Cheikh M. Bamba and Bukula, Andiswa and Mabuya, Rooweither and Dossou, Bonaventure F. P. and Sibanda, Blessing and Buzaaba, Happy and Mukiibi, Jonathan and Kalipe, Godson and Mbaye, Derguene and Taylor, Amelia and Kabore, Fatoumata and Emezue, Chris Chinenye and Aremu, Anuoluwapo and Ogayo, Perez and Gitau, Catherine and Munkoh-Buabeng, Edwin and Memdjokam Koagne, Victoire and Tapo, Allahsera Auguste and Macucwa, Tebogo and Marivate, Vukosi and Mboning, Elvis and Gwadabe, Tajuddeen and Adewumi, Tosin and Ahia, Orevaoghene and Nakatumba-Nabende, Joyce and Mokono, Neo L. and Ezeani, Ignatius and Chukwuneke, Chiamaka and Adeyemi, Mofetoluwa and Hacheme, Gilles Q. and Abdulmumin, Idris and Ogundepo, Odunayo and Yousuf, Oreen and Moteu Ngoli, Tatiana and Klakow, Dietrich. 2022. *MasakhaNER 2.0: Africa-centric Transfer Learning for Named Entity Recognition*. Association for Computational Linguistics.

Alabi, Jesujoba and Adelani, David Ifeoluwa and Mosbach, Marius and Klakow, Dietrich. 2022. *Adapting Pre-trained Language Models to African Languages via Multilingual Adaptive Fine-Tuning*.

Conneau, Alexis and Khandelwal, Kartikay and Goyal, Naman and Chaudhary, Vishrav and Wenzek, Guillaume and Guzmán, Francisco and Grave, Edouard and Ott, Myle and Zettlemoyer, Luke and Stoyanov, Veselin. 2020. *Unsupervised Cross-lingual Representation Learning at Scale*. Association for Computational Linguistics.

Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina. 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Association for Computational Linguistics.

Ogueji, Kelechi and Zhu, Yuxin and Lin, Jimmy. 2021. *Small Data? No Problem! Exploring the Viability of Pretrained Multilingual Language Models for Low-resourced Languages*. Association for Computational Linguistics.

Oladipo, Akintunde. 2024. *Scaling pre-training data and language models for african languages*. University of Waterloo.

Wuraola Fisayo Oyewusi, Olubayo Adekanmbi, Ifeoma Okoh, Vitus Onuigwe, Mary Idera Salami, Opeyemi Osakuade, Sharon Ibejih, and Usman Abdullahi Musa. 2021. Naijaner: Comprehensive named entity recognition for 5 nigerian languages. *arXiv preprint arXiv:2105.00810*.

Pan, Xiaoman and Zhang, Boliang and May, Jonathan and Nothman, Joel and Knight, Kevin and Ji, Heng. 2017. *Cross-lingual Name Tagging and Linking for 282 Languages*. Association for Computational Linguistics.