# SERENGETI: Massively Multilingual Language Models for Africa

**Ife Adebara**[1,⋆] **AbdelRahim Elmadany**[1,⋆] **Muhammad Abdul-Mageed**[1,2] **Alcides Alcoba**[1]
[1]Deep Learning & Natural Language Processing Group, The University of British Columbia
[2]Department of Natural Language Processing & Department of Machine Learning, MBZUAI
{ife.adebara@,a.elmadany@,muhammad.mageed@,alcobaaj@mail.}ubc.ca

## Abstract

Multilingual pretrained language models (mPLMs) acquire valuable, generalizable linguistic information during pretraining and have advanced the state of the art on task-specific finetuning. To date, only $\sim 31$ out of $\sim 2,000$ African languages are covered in existing language models. We ameliorate this limitation by developing SERENGETI, a massively multilingual language model that covers 517 African languages and language varieties. We evaluate our novel models on eight natural language understanding tasks across 20 datasets, comparing to 4 mPLMs that cover $4-23$ African languages. SERENGETI outperforms other models on 11 datasets across the eights tasks, achieving 82.27 average $F_1$. We also perform analyses of errors from our models, which allows us to investigate the influence of language genealogy and linguistic similarity when the models are applied under zero-shot settings. We will publicly release our models for research.[1]

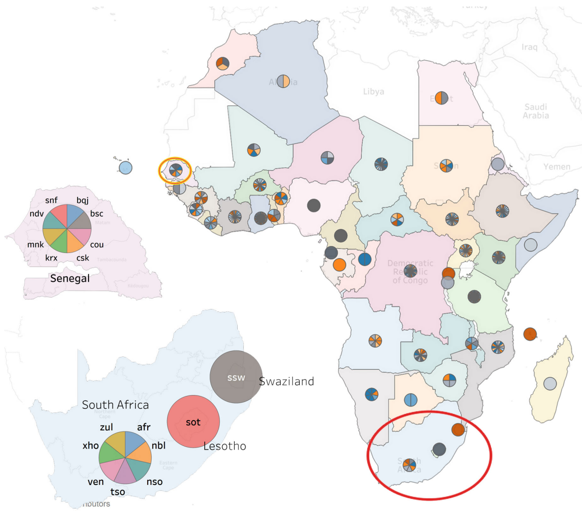



Figure 1: All 517 languages in our dataset across the 50 African countries our data comes from. The language varieties are represented as colored pie shapes within each country. We zero in on South Africa, Lesotho, Swaziland, and Senegal to show detail. We provide a larger map in Appendix A.1.

## 1 Introduction

Pretraining NLP models with a language modeling objective has gained popularity as a precursor to task-specific finetuning (Ettinger, 2020). Pretrained models like BERT (Devlin et al., 2019), ELMo (Peters et al., 2018), Roberta (Liu et al., 2019), GPT (Radford et al., 2018, 2019; Brown et al., 2020a), and BART (Lewis et al., 2020) have advanced the state of the art in a wide variety of tasks, demonstrating how these models acquire valuable, generalizable linguistic information during the pretraining process. However, training language-specific models is possible for only a few languages which have large amounts of data. A popular alternative has been pretrained multilingual language models (mPLM) such as mBERT (Devlin et al., 2019) and XML-R (Conneau et al., 2020). mPLMs are trained on large amounts of unlabelled data from multiple languages so that low resource languages may benefit from shared vocabulary and other linguistic information from high-resource and similar languages in the model. The vast majority of the world's $\sim 7,000$ languages today remain uncovered by mPLMs, however.

African languages are no exception. Although there are few mPLMs that support a small number of African languages (Devlin et al., 2019; Ogueji et al., 2021; Nzeyimana and Niyongabo Rubungo, 2022; Alabi et al., 2022a; Jude Ogundepo et al., 2022; Conneau et al., 2020), these cover only a total of 31 languages. This is grossly inadequate considering that Africa is believed to be home to $\sim 2,000$ languages (Eberhard et al., 2021). Each of these languages encapsulates unique features that are essential in preserving linguistic diversity. The same way every species embodies essential value to the natural ecosystem, each language plays a

[1]https://github.com/UBC-NLP/serengeti
⋆ Authors contributed equally.



*Findings of the Association for Computational Linguistics: ACL 2023*, pages 1498–1537
July 9-14, 2023 

crucial role in the linguistic ecosystem. That is, each language encodes knowledge about people, their traditions, wisdom, and environment, as well as how it is that they interact with the sum of the concepts in their own culture (Adebara and Abdul-Mageed, 2022). This in turn allows people and communities to preserve and transmit their knowledge, values, unique modes of thinking, meaning and expression, history, culture, traditions, and memory to next generations, while participating in society and constructing their future (UNESCO 66260, 2022).

Language technology plays an important role in building inclusive knowledge societies, providing access to education and information, supporting freedom of expression, cultural and linguistic diversity, and further stimulating innovation. This technology thus has great impact on multiple domains, including education, government, health, recreation, among others. This motivates adequate representation of African languages in the ongoing technological revolution. This is also likely to connect Africa to the rest of the world. Building technologies for African languages may also aid languages that may be at risk of falling into a state of disuse at an alarming rate, thus hopefully preventing subsequent language death that may become inevitable (Adebara and Abdul-Mageed, 2022).

Developing LMs that represent a large number of African languages is therefore very crucial for achieving progress in Afrocentric NLP (Adebara and Abdul-Mageed, 2022) and indeed in addressing issues related to representation bias in artificial intelligence and linguistic diversity - two research themes of international relevance (Bender et al., 2021). Motivated by this call for Afrocentric NLP, we introduce **SERENGETI**. SERENGETI is a massively multilingual language model exploiting a large manually-curated dataset for 517 African languages and language varieties. These languages belong to 14 *language families* and are written in 5 different *scripts*. In addition to these African languages, SERENGETI is also pretrained on the top 10 most spoken languages globally.

We also introduce **AfroNLU**, an extensive benchmark exploiting 20 *different datasets* across 28 *different languages and language varieties* for various NLP tasks. For even richer evaluation, we also apply our models to an African language identification task covering all the 517 languages in our pretraining. To the best of our knowledge, AfroNLU is the most extensive and *inclusive* evaluation benchmark proposed to date for African NLP.

Our contributions in this work are as follows: **(1)** we collect a large dataset of 517 African languages and language varieties and exploit it to develop SERENGETI. **(2)** we propose AfroNLU, a new extensive benchmark for African NLU that has the widest and most inclusive coverage for African NLP today. **(3)** we benchmark SERENGETI on AfroNLU and show through meaningful comparisons how our model excels and acquire new SOTA. **(4)** we offer a linguistically motivated analysis of model performance substantiated in language genealogy, allowing us for the first time to derive insights across the widest range of African languages in the African NLP literature to date.

The rest of the paper is organized as follows: In Section 2 we discuss related work. We describe genealogical information in Section 3. Next, we give a detailed description of SERENGETI in Section 4. In Section 5 we describe AfroNLU, the benchmark we create. We present performance of SERENGETI in Section 6 and compare it to other mPLMs. We conclude in Section 7, and outline a number of limitations and use cases for our work in Section 8 and Section 9.

## 2 Related Work

***Afrocentric NLP.*** An *Afrocentric* approach to technology development is crucial for African languages. An afrocentric approach will mean that what technologies to build and how to build, evaluate, and deploy them arises from the needs of local African communities (Adebara and Abdul-Mageed, 2022). We provide more details in Section B in the Appendix.

***African Language Models.*** Here, we briefly describe language models covering any number of African languages. Since we develop encoder-only models in this work, we will restrict our discussion to this category of models. We provide information about the African languages covered by these models in Table 1.

***AfriBERTa*** (Ogueji et al., 2021) is trained using a Transformer with the standard masked language modelling objective and covers 11 African languages. The pretraining corpus for this model is small (only 108.8 million tokens), when compared to many other models. ***AfroLM*** (Dossou et al., 2022) supports 23 African languages, the largest number of African languages before SERENGETI. It is trained on a multi-domain dataset

| Language Model | African languages represented |
|---|---|
| MBERT | Afrikaans, Malagasy, Swahili, Yoruba |
| XLM-R | Afrikaans, Amharic, Hausa, Oromo, Somali, Swahili, Xhosa. |
| KinyarBERT | Kinyarwanda |
| AfriBERTA | Afaan Oromoo, Amharic, Gahuza, Hausa, Igbo, Nigerian Pidgin, Somali, Swahili, Tigrinya and Yoruba |
| Afro-XLMR | Afrikaans, Amharic, Hausa, Igbo, Malagasy, Chichewa, Oromo, Nigerian Pidgin, Kinyarwanda, Kirundi, Shona, Somali, Sesotho, Swahili, isiXhosa, Yoruba, and isiZulu |
| AfroLM | Amharic, Afaan Oromoo, Bambara, Ghomala, Ewe, Fon, Hausa, Igbo, Kinyarwanda, Lingala, Luganada, Luo, Moore, Chewa, Nigerian Pidgin, Shona, Swahili, Setswana, Akan Twi, Wolof, Xhosa, Yoruba, IsiZulu |
| **SERENGETI** | Includes 517 African languages. |

Table 1: Encoder-only models with African languages represented.

from various sources (Adelani et al., 2022a; Alabi et al., 2022b; Jude Ogundepo et al., 2022; Niyongabo et al., 2020). It uses a self-active learning framework and achieves SOTA on NER, sentiment analysis, and text classification. ***Afro-XLM-R*** (Alabi et al., 2022a) uses language adaptation on the 17 most-resourced African languages and three other high-resource foreign languages widely used in Africa (i.e., English, French, and Arabic) simultaneously to provide a single model for cross-lingual transfer learning. Authors show that Afro-XLM-R has competitive results with AfriBERTa and XLM-R on NER, topic classification, and news classification. ***KINYaBERT*** (Nzeyimana and Niyongabo Rubungo, 2022) uses a two-tier BERT architecture that involves a morphological analyzer and explicitly represents morphological information for Kinyawanda–a morphologically rich African language. Authors show that KINYaBERT achieves good convergence and accuracy, and is robust on multiple downstream tasks. ***mBERT*** (Devlin et al., 2019) is a multilingual variant of BERT trained on 104 languages including four African languages. ***XLM-R*** (Conneau et al., 2020) uses a Transformer based architecture and obtains SOTA on cross-lingual classification, sequence labeling, and question answering on 100 languages including eight African languages.

## 3 Genealogy of African Languages

Genealogical or genetic classification groups languages based on their historical and evolutionary relationships. Genetically related languages are often classified into similar families in a hierarchical tree like structure that shows the level of similarity between the languages. Languages with a higher degree of similarity belong to the same class while languages with a lower degree of similarity are further subdivided into different classes and subclasses. Two closely related languages can therefore be viewed as sisters of the same parent language/ancestor–they are languages that evolved over time and/or space from an older parent language (Gerhardt, 2020). Typological classification differs from geneological classification in that the former is based on grammatical features or types (Vossen, 2020). For instance, a typological classification would group tone languages together, or split languages based on their morphological structure into, for instance, isolating or agglutinating languages. Despite this difference, languages that belong to the same family often share similar typological information (Gerhardt, 2020). For example, most Benue-Congo languages are tone languages (Williamson, 2006). In the case of African languages, where typological information is scarcely available (Adebara and Abdul-Mageed, 2022; Güldemann, 2018), utilizing genetic classes may be a useful way to determine typological information. If the typological information of one language in a group is known, we may make a sensible assumption that other languages in that group perhaps share similar features with minor variations. We use geneological classification information in evaluating SERENGETI's behaviour. Specifically, we investigate the relationship between language similarity and model performance in zero-shot scenarios for South African languages in some datasets in our benchmark. We use classification information from Ethnologue (Eberhard et al., 2021) in all our analyses. We provide a broad overview of the families in our models under six broad ancestors in Section D in the Appendix.

## 4 SERENGETI

### 4.1 Pretraining Data

SERENGETI is pretrained using 42GB of data comprising a multi-domain, multi-script collection

| Model | Vocabulary | | #Params | #Lang. (afr/all) | Training Data | | |
|---|---|---|---|---|---|---|---|
| | Tok | Vocab Size | | | Tokens (afr/all) | Size (afr/all) | Source |
| xlmr | SP | 250k | 270M | 8 / 100 | UNK/164B | UNK/2.4 GB | CC-100 |
| mbert | WP | 110K | 110M | 4 / 100 | UNK/12.8B | UNK/100GB | Books, Wiki. |
| Afro-XLMR | SP | 70.6K | 270M | 17 / 20 | — | 21.6 GB | mC4, CC, BBC, VOA |
| AfriBERTa | WP | 70k | 111M | 11 / 11 | 108.8M | 0.94 GB | BBC, CC |
| AfroLM | SP | 250K | 264M | 23/23 | — | 0.73GB | mC4, CC, BBC, VOA |
| SERENGETI-E110 | WP | 110K | 170M | 517 / 527 | 7.1B/8.6B | 40/42GB | RT, News, GD, HD, EC |
| SERENGETI-E250 | WP | 250K | 277M | 517/527 | 7.1B/8.6B | 40/42GB | RT, News, GD, HD, EC |
| SERENGETI | SP | 250K | 278M | 517/527 | 7.1B/8.6B | 40/42GB | RT, News, GD, HD, EC |

Table 2: Models with African languages that we compare SERENGETI with. SP: SentencePiece, WP: WordPiece. Data sources include - CC: CommonCrawl, EC: Existing corpora, GD: Government documents, HD: Health documents, RT: Religious text, UNK: Unknown.

of texts that we manually curate. The pretraining data covers 517 African languages and the 10 most spoken languages globally (i.e., Arabic, English, French, German, Greek, Italian, Portuguese, Russian, Spanish, and Turkish). The multi-domain dataset comprises texts from religious, news, government documents, health documents, and existing corpora written in five scripts from the set *{Arabic, Coptic, Ethiopic, Latin, and Vai}*. For the top ten foreign languages, we randomly select 1M paragraphs from Wikipedia for each language to use in our overall pretraining data. We provide further details of the pretraining data in Section C in the Appendix. We also show all languages in our pretraining data in Tables F.1, F.2, and F.3.

### 4.2 Preprocessing

To prepare the raw data for pretraining, we perform light preprocessing to retain a faithful representation of the naturally occurring text. Specifically, we ensure that images and non-text materials are not in our dataset by using regular expression and manual curation techniques. We do not perform any further preprocessing of the data before splitting the text off into tokens. For tokenization, we use a WordPiece tokenizer (Song et al., 2021). We experiment with two vocabulary sizes, 110K and 250K.

### 4.3 SERENGETI Models

We pretrain both Electra style (Clark et al., 2020b; Chi et al., 2021) as well as XLM-R style (Conneau et al., 2020) models, as follows.

**SERENGETI-E110 and SERENGETI-E250.** We first pretrain Electra (Chi et al., 2021) style models. Electra uses a multilingual replaced token detection (MRTD) objective for training. Unlike other training objectives, the goal of MRTD is to distinguish real input tokens from corrupted tokens. Models built with this objective are pretrained as discriminators rather than generators. We train the models with two vocabulary sizes, 110K and 250K, and hence refer to them as SERENGETI-E110 and SERENGETI-E250. Each of these models has 12 layers and 12 attention heads. We pretrain each model for 40 epochs with a sequence length of 512, a learning rate of $2e-4$ and a batch size of 216 and 104 for the SERENGETI-E110 and SERENGETI-E250, respectively. We pre-train the models on 1 Google Cloud TPU with 8 cores (v3.8) from TensorFlow Research Cloud (TFRC). [2]

**SERENGETI Model.** Apart form the Electra models, we also experiment with an XLM-R base architecture. We train the model with a 250K vocabulary size for 20 epochs. This model has 12 layers and 12 attention heads, a sequence length of 512 and a batch size of 8. We pre-train this model on 80 M50 AMD Pod GPUs with 16G ram. Our XLM-R model has better performance compared to the Electra models as we will show. We provide information about each model we build and compare with in Table 2.

## 5 AfroNLU Benchmark

Our goal is to evaluate our models extensively, and so we combine all available datasets we could acquire to create an evaluation benchmark that we refer to as **AfroNLU**. AfroNLU is composed of *seven different tasks*, covering both token and sentence level tasks, across 18 different datasets. The benchmark covers a total of 32 *different languages and language varieties*. In addition we evaluate our best model (SERENGETI) on an African lan-

[2] https://www.tensorflow.org/tfrc.

| Cluster | Dataset | Languages | TRAIN | DEV | TEST |
|---|---|---|---|---|---|
| NER | masakaner-v1⋆ | amh, hau, ibo, kin, lug, luo, pcm, swh, wol, yor | 443,692 | 60,515 | 134,126 |
| | masakaner-v2⋆ | bam, bbj, ewe, fon, hau, ibo, kin, lug, mos, nya, pcm, sna, swa, tsn, twi, wol, xho, yor, zul | 2,537,792 | 362,837 | 726,830 |
| | masakaner-east⋆ | amh, kin, lug, luo, swh | 162,388 | 21,206 | 46,407 |
| | masakaner-eastwest⋆ | amh, hau, ibo, kin, lug, luo, pcm, swh, wol, yor | 416,113 | 56,512 | 126,176 |
| | masakaner-west⋆ | hau, ibo, pcm, wol, yor | 253,725 | 35,306 | 79,769 |
| | nchlt-ner ⋆ | afr, nbl, nso, sot, ssw, tsn, tso, ven, xho, zul | 1,749,372 | 219,703 | 215,616 |
| | yoruba-twi-ner⋆ | yor | 20,237 | 2,811 | 5,471 |
| | wikiann⋆ | afr, amh, ibo, mlg, kin, som, swh, yor | 9,244 | 9,240 | 9,424 |
| Phrase Chunking | phrase-chunk⋆ | afr, nso, sot, ssw, tsn, tso, ven, zul | 107,492 | 12,972 | 13,389 |
| POS | igbo-pos⋆ | ibo | 756,775 | 94,692 | 95,048 |
| News | amharic-news† | amh | 41,185 | 5,148 | 5,149 |
| | kinnews† | kir | 15,308 | 1,701 | 4,254 |
| | kirnews† | run | 3,320 | 369 | 923 |
| | swahili-news-v0.2† | swh | 19,986 | 2,221 | 7,338 |
| Sentiment Analysis | bambara-v2† | bam | 2,436 | 305 | 305 |
| | pidgin-tweet† | pcm | 11,200 | 1,400 | 1,400 |
| | yosm† | yor | 800 | 200 | 500 |
| Topic | hausa-topic† | hau | 2,045 | 290 | 582 |
| | yoruba-topic† | yor | 1,340 | 189 | 379 |
| QA | qa-swahili† | swh | 49,881 | 5,077 | 499 |
| LID | AfroLID† | 517 African Languages | 2,496,980 | 25,850 | 51,400 |
| | Afri-Senti | amh, hau, ibo, pcm, swh, yor | | | - |

Table 3: Distribution of AfroNLU datasets. ⋆ indicates that datasize is measured at token level. † indicates data size measured at sentence level.

| Tasks | AfriBERTa | Afro-XLMR | KinyaBERT | SERENGETI |
|---|---|---|---|---|
| NER | ✓ | ✓ | ✓ | ✓ |
| PC | — | — | — | ✓ |
| POS | — | — | — | ✓ |
| NC | ✓ | ✓ | — | ✓ |
| SA | — | ✓ | — | ✓ |
| TC | — | ✓ | ✓ | ✓ |
| QA | — | — | — | ✓ |
| LID | — | — | — | ✓ |
| GLUE | — | — | ✓ | — |

Table 4: Tasks evaluation comparison across different African language MLMs. NER: named entity recognition, PC: phrase chunking, POS: part of speech, NC: news classification, SA: sentiment analysis, TC: topic classification, QA: question answering, LID: language identification.

guage identification (LID) task covering all the 517 languages in our pretraining collection. For LID, we use two datasets to test SERENGETI. This puts AfroNLU at a total of *20 different datasets* and eight different tasks. To the best of our knowledge, our evaluation benchmark is the most extensive compared to previous published research. We provide detailed statistics of the datasets comprising AfroNLU in Table 3. We also provide a detailed comparison of our AfroNLU benchmark with evaluation data from other models in Table 4.

We now describe each of the downstream tasks in AfroNLU.

### 5.1 Named Entity Recognition (NER)

We evaluate our models on NER datasets across multiple languages. We use MasakhaNER data (Ifeoluwa Adelani et al., 2021), WikiAnn (Pan et al., 2017; Rahimi et al., 2019), Yoruba-Twi NER data (Alabi et al., 2020), Distance Supervision NER (DS NER) Data (Hedderich et al., 2020) and multiple NER data from SADiLaR. For our experiments, we use the region aggregates on MasakhaNER. Specifically, we use MasakhaNER-east, MasakhaNER-west, and MasakhaNER-eastwest. MasakhaNER-east includes NER data for Amharic, Kinyawanda, Luganda, Luo, and Swahili. MasakhaNER-west includes NER data for Hausa, Igbo, Nigerian-Pidgin, Wolof, and Yoruba. MasakhaNER-eastwest, on the other hand, includes a combination of MasakhaneNER-east and MasakhaneNER-west. Data from SADiLaR cover ten indigenous South African languages and is annotated for person, organisation, location, and miscellaneous named entities. Miscellaneous named entities refer to all rigid designators that do not fall into one of the other categories, including temporal expressions (dates and times), URLs, numerical expressions, publications, names of languages, nationalities, among others. More details about the datasets are in Table 3.

### 5.2 Part of Speech Tagging

We test our models on POS tagging datasets for Igbo taken from IgboNLP (Onyenwe et al., 2018, 2019). In Table 3, we provide the statistical details for the dataset.

### 5.3 Phrase Chunks

We evaluate our models on phrase chunks datasets for ten Indigenous languages of South Africa (see Table 3). The data has annotations for noun, verb, adjective, adverbial, and prepositional phrase chunks. Words not belonging to these phrase types are labelled with the tag *O*.

### 5.4 Sentiment Analysis

We finetune our model on three sentiment analysis datasets, including Bambara Sentiment dataset (Diallo et al., 2021), YOSM–a new Yorùbá Sentiment Corpus for Movie Reviews (Shode et al., 2022), and the Nigerian Pidgin sentiment dataset (Oyewusi et al., 2020), respectively. Some details of these datasets is in Table 3.

### 5.5 News classification

We use news classification datasets for Amharic (Azime and Mohammed, 2021), Kinyarwanda (Niyongabo et al., 2020), Kirundi (Niyongabo et al., 2020), and Swahili (David, 2020a,b). The Amharic dataset contains six classes–news, sport, politics, international news, business, and entertainment. The Swahili dataset also has six categories including local news, international, finance, health, sports, and entertainment. The datasets for Kinyarwanda and Kirundi have 14 and 12 categories each, respectively. Again, data statistics are in Table 3.

### 5.6 Topic classification

We include topic classification datasets for Yorùbá and Hausa (Hedderich et al., 2020). The Yorùbá and Hausa datasets contain news titles collected from VOA Hausa and BBC Yorùbá news sites. The Yorùbá dataset has seven topics–Nigeria, Africa, world, entertainment, health, sports, and politics, while the Hausa dataset is categorized into five topics - Nigeria, Africa, world, health, and politics. In Table 3, we provide details about the data split sizes.

### 5.7 Question Answering

We use TYDIA question answering dataset (Clark et al., 2020a). The dataset has a primary task and a gold passage task. The primary task has two subtasks, one for passage selection and another that is a minimal answer span. For the passage selection subtask, a list of passages is given and the required response is either the index of the passage where the answer to the question is or null (if no answer exists in the passage). The minimal answer span subtask on the other hand gives a full article and the expected answer is either the start and end byte indices of the minimal span that answers the question, yes or no response, or null (if no minimal answer exists). For the gold passage task, a correct answer is predicted from a passage containing one answer. This is similar to existing reading comprehension. We use the Kiswahili dataset alone, since it is the only African language in the dataset. Details about the data splits can be found in Table 3.

### 5.8 Language Identification

We also evaluate SERENGETI on the task of language identification (LID). LID focuses on identifying the human language a piece of text or speech segment belongs to, making automatic LID an important first step in processing human language appropriately (Tjandra et al., 2021; Thara and Poornachandran, 2021). We use datasets from AfroLID (Adebara et al., 2022b) for this task. AfroLID data is a multi-genre, multi-script dataset for 517 African languages. We compare the performance of AfroLID data on our models with performance on AfroLID tool. To ensure a fair comparison, the data used for AfroLID is completely different from the data used for SERENGETI. We also evaluate our LID model on AfriSenti dataset (Muhammad et al., 2022; Yimam et al., 2020).

## 6 Experimental Setup and Evaluation

We evaluate SERENGETI on eight task clusters in the benchmark, and report results on our Test set in Table 5. We also report performance on our Dev set in Table E.1 (Appendix). For each task cluster, we finetune for a maximum of 25 epochs with a patience value of five. We compare results from SERENGETI, SERENGETI-E110, and SERENGETI-E250 to encoder-only models covering any number of African languages. Specifically, we compare with XLMR, mBERT, Afro-XLMR, and AfriBERTa. We report the results of each experiment as an average of three runs, showing

| Cluster | Dataset | SOTA | XLMR | mBERT | Afro-XLMR | AfriBERTa | SERENGETI-E110 | SERENGETI-E250 | SERENGETI |
|---|---|---|---|---|---|---|---|---|---|
| NER | masakaner-v1 | $84.80^{\pm 0.3}$‡ ‡ ‡ | $81.41^{\pm 0.26}$ | $78.57^{\pm 0.53}$ | $84.16^{\pm 0.45}$ | $81.42^{\pm 0.30}$ | $81.23^{\pm 0.32}$ | $81.54^{\pm 0.68}$ | **84.53** $^{\pm 0.56}$ |
| | masakaner-v2 | $87.00^{\pm 1.2}$‡ ‡ ‡ | $87.17^{\pm 0.18}$ | $84.82^{\pm 0.96}$ | $88.69^{\pm 0.12}$ | $86.22^{\pm 0.06}$ | $86.57^{\pm 0.27}$ | $86.69^{\pm 0.29}$ | **88.86** $^{\pm 0.25}$ |
| | masakaner-east | 80.62⋆ | $80.38^{\pm 0.56}$ | $78.33^{\pm 1.25}$ | $83.02^{\pm 0.31}$ | $79.31^{\pm 0.92}$ | $80.53^{\pm 0.71}$ | $81.26^{\pm 0.68}$ | **83.75** $^{\pm 0.26}$ |
| | masakaner-eastwest | 82.34⋆ | $82.85^{\pm 0.38}$ | $82.37^{\pm 0.90}$ | **86.31** $^{\pm 0.30}$ | $82.98^{\pm 0.44}$ | $82.90^{\pm 0.49}$ | $83.67^{\pm 0.44}$ | $85.94^{\pm 0.27}$ |
| | masakaner-west | 83.11⋆ | $82.85^{\pm 0.79}$ | $83.99^{\pm 0.39}$ | **86.78** $^{\pm 0.44}$ | $84.08^{\pm 0.32}$ | $82.06^{\pm 0.67}$ | $83.45^{\pm 0.81}$ | $86.27^{\pm 0.94}$ |
| | nchlt-ner | — | $71.41^{\pm 0.07}$ | $70.58^{\pm 0.26}$ | $72.27^{\pm 0.14}$ | $68.74^{\pm 0.29}$ | $64.46^{\pm 0.37}$ | $64.42^{\pm 0.24}$ | **73.18** $^{\pm 0.24}$ |
| | yoruba-twi-ner | — | $61.18^{\pm 2.19}$ | $70.37^{\pm 0.61}$ | $58.48^{\pm 1.85}$ | $69.24^{\pm 3.05}$ | $61.77^{\pm 1.24}$ | $57.99^{\pm 2.61}$ | **71.25** $^{\pm 1.73}$ |
| | wikiann | | $83.82^{\pm 0.39}$ | $82.65^{\pm 0.77}$ | **86.01** $^{\pm 0.83}$ | $83.05^{\pm 0.20}$ | $83.17^{\pm 0.54}$ | $84.85^{\pm 0.53}$ | $85.83^{\pm 0.94}$ |
| Phrase Chunking | phrase-chunk | — | $88.86^{\pm 0.18}$ | $88.65^{\pm 0.06}$ | $90.12^{\pm 0.12}$ | $87.86^{\pm 0.20}$ | $90.39^{\pm 0.21}$ | $89.93^{\pm 0.33}$ | **90.51** $^{\pm 0.04}$ |
| POS | igbo-pos | — | $85.50^{\pm 0.08}$ | $85.42^{\pm 0.13}$ | $85.39^{\pm 0.21}$ | $85.43^{\pm 0.05}$ | $85.50^{\pm 0.16}$ | **85.61** $^{\pm 0.13}$ | $85.54^{\pm 0.08}$ |
| News Classification | amharic-news | — | $84.97^{\pm 0.55}$ | $59.01^{\pm 1.47}$ | $86.18^{\pm 0.85}$ | $86.54^{\pm 1.20}$ | $86.50^{\pm 0.71}$ | $86.34^{\pm 0.30}$ | **86.82** $^{\pm 0.72}$ |
| | kinnews | | $76.58^{\pm 0.70}$ | $77.45^{\pm 0.43}$ | $79.13^{\pm 0.53}$ | $80.40^{\pm 1.50}$ | **81.43** $^{\pm 1.02}$ | $80.38^{\pm 1.36}$ | $79.80^{\pm 0.68}$ |
| | kirnews | — | $57.18^{\pm 3.44}$ | $74.71^{\pm 2.56}$ | $87.67^{\pm 0.92}$ | **89.59** $^{\pm 0.27}$ | $78.75^{\pm 3.24}$ | $86.60^{\pm 1.28}$ | $87.53^{\pm 2.31}$ |
| | swahili-news-v0.2 | — | $87.50^{\pm 0.91}$ | $85.12^{\pm 0.93}$ | $87.49^{\pm 1.26}$ | $87.91^{\pm 0.36}$ | $87.33^{\pm 0.28}$ | $86.12^{\pm 1.30}$ | **88.24** $^{\pm 0.99}$ |
| Sentiment Analysis | bambara-v2 | 64.00† | $47.17^{\pm 1.83}$ | $64.56^{\pm 1.71}$ | $59.40^{\pm 0.56}$ | $65.06^{\pm 2.08}$ | $65.07^{\pm 2.59}$ | **65.76** $^{\pm 2.02}$ | $63.36^{\pm 3.31}$ |
| | pidgin-tweet | — | $70.42^{\pm 0.68}$ | $68.59^{\pm 0.47}$ | **71.40** $^{\pm 0.51}$ | $69.19^{\pm 0.97}$ | $71.06^{\pm 0.39}$ | $70.46^{\pm 1.02}$ | $69.74^{\pm 0.92}$ |
| | yosm | 87.20‡ | $85.57^{\pm 1.09}$ | $85.25^{\pm 0.25}$ | $87.46^{\pm 0.42}$ | **88.66** $^{\pm 0.23}$ | $86.86^{\pm 0.95}$ | $85.58^{\pm 1.51}$ | $87.86^{\pm 0.81}$ |
| Topic | hausa-topic | 48.52 †† | $85.80^{\pm 1.45}$ | $81.38^{\pm 0.42}$ | $88.67^{\pm 0.30}$ | **92.59** $^{\pm 0.69}$ | $88.52^{\pm 1.31}$ | $89.07^{\pm 0.95}$ | $89.93^{\pm 0.49}$ |
| | yoruba-topic | 54.93 †† | $54.69^{\pm 2.89}$ | $71.79^{\pm 1.43}$ | $75.13^{\pm 1.40}$ | **81.79** $^{\pm 0.66}$ | $65.22^{\pm 4.72}$ | $66.34^{\pm 4.09}$ | $79.87^{\pm 1.61}$ |
| QA | qa-swahili | 81.90 ‡‡ | $82.79^{\pm 1.93}$ | **83.40** $^{\pm 0.78}$ | $79.94^{\pm 0.39}$ | $57.3^{\pm 1.8}$ | $79.76^{\pm 0.52}$ | $81.25^{\pm 1.33}$ | $80.01^{\pm 0.78}$ |
| | **AfroNLU Score** | | 76.91 | 77.85 | 81.09 | 80.37 | 79.45 | 79.87 | **82.44** |

Table 5: Performance of models on seven AfroNLU benchmark TEST datasets. ($F_1$) score is the evaluation metric. Our model (SERENGETI) significantly outperforms AfriBERTa (the 2[nd] in row) on 13/18 datasets and achieve SOTA on 9/18 datasets. **SOTA** as reported on ⋆(Ifeoluwa Adelani et al., 2021), †(Diallo et al., 2021), ‡(Shode et al., 2022), ††(Hedderich et al., 2020) and ‡‡(Clark et al., 2020a), ‡ ‡ ‡(Adelani et al., 2022b). We use a dash (-) to represent tasks without a known SOTA.

the standard deviation. We also evaluate SERENGETI on language identification and show results on Afrolid in Table 6 and on Afrisenti in Table 7. For multilingual datasets in each task, we show evaluation results per language, comparing the performance of various models in Table E.4 in the Appendix.

| Task | AfroLID | SERENGETI |
|---|---|---|
| Dev | 96.14⋆ | **97.64** $^{\pm 0.02}$ |
| Test | 95.95⋆ | **97.41** $^{\pm 0.02}$ |

Table 6: Performance of SERENGETI on African LID ($F_1$). ⋆ Results as reported in Adebara et al. (2022b).

| | AfroLID | SERENGETI |
|---|---|---|
| Amharic (amh) | 97.00 | **99.50** $^{\pm 0.01}$ |
| Hausa (hau) | 89.00 | **98.09** $^{\pm 0.02}$ |
| Igbo (ibo) | 46.00 | **95.28** $^{\pm 0.00}$ |
| Nigerian Pidgin (pcm) | 56.00 | **77.73** $^{\pm 0.01}$ |
| Swahili (swh) | 96.00 | **98.66** $^{\pm 0.02}$ |
| Yoruba (yor) | 82.00 | **98.96** $^{\pm 0.00}$ |

Table 7: Comparison between AfroLID (Adebara et al., 2022b) and SERENGETIon AfriSenti Dev dataset.

### 6.1 Performance Analysis

We report the results for seven of our eight tasks in Table 5.

**Named Entity Recognition (NER).** SERENGETI sets a new SOTA on six out of eight datasets on the NER cluster. The lowest $F_1$ across all models are on NCHLT and Yoruba-Twi datasets (on both Dev and Test). SERENGETI achieves best performance on both of these datasets on Test (with 73.18 $F_1$ on the first and 71.25 on the second).

**Phrase Chunking.** SERENGETI outperforms all models on the phrase chunking task on both Dev and Test data, reaching 90.51 $F_1$ on Test.

**Part of Speech (POS) Tagging.** In the POS tagging task, SERENGETI outperformed all other models in the Dev. and Test sets.

**News Classification.** Our SERENGETI outperforms other models on three out of four datasets on Test data (and on two datasets on Dev).[3]. We do not report SOTA results for Amharic, Kirnews, and Kinnews datasets because their authors report performance in accuracy (and so are not comparable to our results). We show performance of SERENGETI on each category in the news classification cluster in Figure E.1 in the Appendix.

**Sentiment Analysis.** SERENGETI-E250 outperforms other models on one out of three tasks in our sentiment analysis task cluster. Afro-XMLR and AfriBERTa outperform other models on one each. To further investigate performance, we conduct an error analysis on the three sentiment datasets (see Figure E.2 in the Appendix).

**Topic Classification.** AfriBERTa outperforms

[3]Our SERENGETI-E110 outperforms SERENGETI on one dataset in Dev and Test sets

other models on both tasks in our topic classification cluster, followed by SERENGETI. We show confusion matrices for Hausa and Yoruba topic classification in Figure E.3 in the Appendix.

**Language Identification.** SERENGETI outperforms AfroLID on AfroLID and AfriSenti data (see Table 6 and 7 for details). We also compare the performance of SERENGETI to AfroLID, and Franc[4], on the 88 African languages represented in Franc in Table E.3 (Appendix). SERENGETI outperforms AfroLID and Franc with an average $F_1$ score of 96.29. SERENGETI outperforms both models on 59 languages and has similar results with AfroLID on 19 languages. Next, we evaluate the performance of SERENGETI on Creole languages. Again, we record improvement in results for Creole languages when compared with AfroLID. SERENGETI outperforms AfroLID in 7 out of 9 languages and acquires similar scores on 2 languages. We assume that the addition of the ten most spoken languages to the pretraining data for SERENGETI may have helped the model learn the Creoles better. This is because Creoles share some features including vocabularies and syntax with some of those top ten languages.

### 6.2 Error Analysis

In the sentiment analysis cluster, best performance is recorded for positive categories while negative categories have the worst performance. A fine-grained analysis of the Yoruba sentiment dataset found that SERENGETI failed to correctly categorize sentiment if the polarity item(s) were not seen in training, can be associated with both positive and negative sentiments, the polarity item(s) is a negation, or if ambivalent markers are present in the sentence. We provide a table showing examples of each type of error we found in Table E.2 in the Appendix. For the news classification task, politics and tourism are the best performing classes while education and relationships have the worst performance on kirnews and kinnews respectively. It is important to mention that the worst performing categories do not have the smallest data sizes. For the topic classification, the best performance is on the world class for Hausa topic modelling while entertainment and sport have best performance for Yoruba. The worst performance is on Nigeria and health for Hausa and Yoruba topic datasets respectively.

[4]A publicly available LID tool covering 88 African languages.

### 6.3 Imbalanced Distribution

We find imbalances in the class distributions for all datasets except YOSM. We find a positive correlation between the size of each category in a dataset and the model accuracy. We also find a positive correlation with the number of examples in a specific class and the accuracy we acquire. We provide confusion matrices that represents the sizes of each category and the performance of SERENGETI in Figures E.4, E.5, and E.6 in the Appendix.

### 6.4 Genealogy & Language Contact

Our preliminary analyses show that language similarity may improve model performance in zero-shot settings. This we believe is due to high cross-lingual transfer information (Conneau et al., 2020) from similar languages. Similar languages often share many features (e.g., vocabulary, syntax, and script) sometimes up to a point of mutual intelligibility (Nassenstein, 2019; Arndt, 2015; Roy-Campbell, 2006). Languages in contact may also have such similarities. By *language in contact*, we mean all languages that speakers of a specific language interact with and influence. A language can be in contact with another due to trade, geographic proximity, migration, or even colonization. Languages in contact can influence each other in multiple ways, such as borrowing words, grammatical structures, phonology, or orthographic conventions (Matras, 2009). To illustrate our hypothesis, we select two datasets with South African (SA) languages in AfroNLU - NCHLT-ner and phrase-chunk. We select SA languages because they are contact languages (see Figure D.5 in Appendix for a genealogical classification tree that highlights the SA languages.) (Nassenstein, 2019; Arndt, 2015; Roy-Campbell, 2006).

To determine the significance of language similarity and language contact in our own zero-shot settings, we measure the Jaccard similarity between the pretraining data for the SA languages (see Table 8). We find strong similarities between some of these languages (see bolded examples in Table 8). We also finetune a BERT model and compare the performance of BERT with MBERT. We do this because BERT does not include any similar language in its representation.

XLM-R, mBERT, and AfriBERTa are not trained on most SA languages but have high scores in zero-shot settings see Table 9 and Table E.4 in Appendix. We argue that XLM-R in addition to

| | afr | nbl | nso | sot | ssw | tsn | tso | ven | xho | zul | kin | lug | nya | run | sna | som |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **afr** | 1 | 0.28 | 0.35 | 0.26 | 0.27 | 0.36 | 0.29 | 0.22 | **0.42** | 0.38 | 0.34 | 0.38 | 0.26 | 0.25 | 0.25 | **0.43** |
| **nbl** | 0.28 | 1 | **0.47** | **0.41** | **0.62** | 0.26 | **0.48** | **0.42** | **0.41** | **0.55** | 0.37 | 0.35 | **0.48** | **0.43** | **0.46** | 0.35 |
| **nso** | 0.35 | **0.47** | 1 | **0.55** | **0.47** | 0.38 | **0.51** | **0.40** | **0.42** | **0.50** | **0.40** | 0.38 | **0.42** | 0.39 | 0.39 | **0.42** |
| **sot** | 0.26 | **0.41** | **0.55** | 1 | **0.43** | 0.27 | **0.52** | **0.46** | 0.31 | **0.41** | 0.33 | 0.29 | **0.45** | **0.40** | 0.39 | 0.34 |
| **ssw** | 0.27 | **0.62** | **0.47** | **0.43** | 1 | 0.25 | **0.50** | **0.44** | 0.38 | **0.52** | 0.36 | 0.33 | **0.48** | **0.43** | **0.43** | 0.34 |
| **tsn** | 0.36 | 0.26 | 0.38 | 0.27 | 0.25 | 1 | 0.28 | 0.21 | 0.39 | 0.36 | 0.31 | 0.36 | 0.25 | 0.24 | 0.23 | 0.37 |
| **tso** | 0.29 | **0.48** | **0.48** | **0.52** | **0.50** | 0.28 | 1 | **0.47** | 0.37 | **0.48** | 0.38 | 0.34 | **0.51** | **0.44** | **0.44** | 0.37 |
| **ven** | 0.22 | **0.42** | **0.40** | **0.46** | **0.44** | 0.21 | 0.47 | 1 | 0.27 | 0.35 | 0.29 | 0.26 | **0.44** | 0.38 | **0.41** | 0.29 |
| **xho** | **0.42** | **0.41** | **0.42** | 0.31 | 0.38 | 0.39 | 0.37 | 0.27 | 1 | **0.56** | **0.41** | **0.47** | 0.35 | 0.33 | 0.32 | **0.45** |
| **zul** | 0.38 | **0.55** | **0.50** | **0.41** | **0.52** | 0.36 | **0.48** | 0.35 | **0.56** | 1 | **0.44** | **0.44** | **0.44** | **0.40** | 0.39 | **0.45** |

Table 8: Jaccard Similarity for South African languages and some languages that are genealogically similar to them. Each of the 10 South African languages are represented on each row. The genealogically similar languages we explore are after the horizontal lines. Specifically, we have: Kinyarwanda (kin), Luganda (lug), Chichewa (nya), Rundi (run), Shona (sna) and Somali (som). We highlight similarity scores of 0.4 and above in bold face.

| Dataset | Lang | XLMR | BERT | mBERT | Affro-XLMR | AfriBERTa | SERENGETI |
|---|---|---|---|---|---|---|---|
| NCHLT-NER | afr | $80.68^{\pm0.75}$ | 71.47 | $80.08^{\pm0.29}$ | $80.55^{\pm0.11}$ | $74.5^{\pm0.64}$ | $\mathbf{81.57}^{\pm0.59}$ |
| | nbl | $74.64^{\pm0.66}$ | 61.02 | $73.48^{\pm0.18}$ | $75.26^{\pm0.28}$ | $72.28^{\pm0.67}$ | $\mathbf{77.13}^{\pm0.67}$ |
| | nso | $77.0^{\pm1.23}$ | 64.27 | $78.75^{\pm0.45}$ | $80.13^{\pm0.51}$ | $75.45^{\pm1.09}$ | $\mathbf{80.69}^{\pm0.64}$ |
| | sot | $54.71^{\pm1.51}$ | 49.75 | $54.68^{\pm0.49}$ | $55.57^{\pm0.2}$ | $54.09^{\pm0.98}$ | $\mathbf{56.26}^{\pm1.52}$ |
| | ssw | $71.75^{\pm0.65}$ | 65.18 | $71.24^{\pm0.75}$ | $72.35^{\pm1.02}$ | $69.38^{\pm0.58}$ | $\mathbf{73.37}^{\pm0.82}$ |
| | tsn | $77.02^{\pm0.22}$ | 70.96 | $76.35^{\pm0.47}$ | $77.68^{\pm0.96}$ | $73.89^{\pm1.41}$ | $\mathbf{79.05}^{\pm0.75}$ |
| | tso | $74.24^{\pm0.08}$ | 65.09 | $72.95^{\pm0.67}$ | $74.85^{\pm0.43}$ | $71.05^{\pm0.9}$ | $\mathbf{75.13}^{\pm0.31}$ |
| | ven | $64.06^{\pm0.31}$ | 61.51 | $63.11^{\pm1.27}$ | $64.39^{\pm0.36}$ | $63.24^{\pm1.26}$ | $\mathbf{65.42}^{\pm0.76}$ |
| | xho | $70.77^{\pm2.45}$ | 58.17 | $68.54^{\pm1.44}$ | $72.37^{\pm0.39}$ | $67.00^{\pm1.27}$ | $\mathbf{72.92}^{\pm0.29}$ |
| | zul | $69.44^{\pm0.62}$ | 54.27 | $67.74^{\pm1.46}$ | $70.28^{\pm0.49}$ | $67.17^{\pm0.15}$ | $\mathbf{71.20}^{\pm0.44}$ |
| Phrase Chunk | afr | $95.34^{\pm0.16}$ | 89.92 | $95.68^{\pm0.30}$ | $95.13^{\pm0.06}$ | $90.22^{\pm0.81}$ | $\mathbf{96.01}^{\pm0.14}$ |
| | nso | $96.57^{\pm0.61}$ | 95.26 | $96.85^{\pm0.55}$ | $\mathbf{98.36}^{\pm0.2}$ | $96.47^{\pm0.14}$ | $98.28^{\pm0.1}$ |
| | sot | $82.93^{\pm0.38}$ | 80.59 | $83.08^{\pm0.78}$ | $85.28^{\pm0.61}$ | $82.18^{\pm0.93}$ | $\mathbf{85.69}^{\pm0.76}$ |
| | ssw | $82.9^{\pm1.03}$ | 82.09 | $81.91^{\pm0.47}$ | $\mathbf{84.73}^{\pm0.18}$ | $83.24^{\pm0.11}$ | $83.45^{\pm0.12}$ |
| | tsn | $92.77^{\pm0.16}$ | 92.09 | $92.64^{\pm0.66}$ | $94.11^{\pm0.49}$ | $92.71^{\pm0.42}$ | $\mathbf{94.03}^{\pm0.19}$ |
| | tso | $86.42^{\pm0.46}$ | 86.75 | $86.90^{\pm0.31}$ | $87.39^{\pm0.18}$ | $86.73^{\pm0.95}$ | $\mathbf{89.32}^{\pm0.43}$ |
| | ven | $92.31^{\pm0.45}$ | 92.32 | $90.47^{\pm0.32}$ | $92.42^{\pm0.68}$ | $92.02^{\pm0.33}$ | $\mathbf{92.54}^{\pm0.21}$ |
| | zul | $87.30^{\pm0.26}$ | 84.93 | $87.29^{\pm1.04}$ | $88.67^{\pm0.66}$ | $85.74^{\pm0.55}$ | $\mathbf{90.05}^{\pm0.81}$ |

Table 9: Performance of mPLMs and BERT on each language in NCHLT-NER and Phrase-Chunk datasets we use for the genealogy analysis. ($F_1$) score is the evaluation metric. We use **Red** highlights to indicate languages in zero-shot setting. We evaluate BERT, a monolingual model as a sanity check for our evaluation.

cross-lingual transfers from other languages acquires representation from afr and xho where xho alone shares more than 0.4 similarity with afr, nbl, nso, and zul. mBERT also learns representation from afr while AfriBERTa learns representations from Gahuza which is a code-mixed variety of KIN and RUN. BERT on the other hand significantly performs lower than MBERT in all languages except on ssw, and ven (Phrase chunk). SERENGETI, however, outperforms other models on these languages which demonstrates the impact of pretraining on each of these languages.

These analyses are in no way conclusive, but do provide insights on how linguistic information may impact model performance in zero-shot settings. Future work can further probe the influence of similar languages in a more in-depth fashion. (See Appendix F for detailed analysis).

## 7 Conclusion

We reported our efforts to develop SERENGETI, a suite of three massively multilingual language models for African NLP. SERENGETI outperforms 4 mPLMs on 11 datasets across 8 tasks. We provide extensive evaluations of model outputs, including zero-shot performance of the mPLMs. We also offer broad linguistically-motivated analyses of model performance.

## 8 Limitations

We identify the following limitations for our work:

1. Due to limited access to a wide network of native speakers from the majority of languages, we were able to manually inspect only a subset of languages present in our pretraining data. Specifically, we could only manually evaluate Afrikaans, Yorùbá, Igbo, Hausa, Luganda, Kinyarwanda, Chichewa, Shona, Somali, Swahili, Xhosa, Bemba, and Zulu. Future work should focus on increasing the subset of languages evaluated manually in order to ensure quality. We believe automatic analyses are not sufficient before development of models that get deployed in particular applications.

2. Another limitation is related to our inability to perform extensive analysis of biases and hateful speech present in our pretraining data. Again, this is due to relatively restricted access to native speakers (and even automated tools) to perform this analysis. As a result, we cannot fully ensure that our models is free from biases and socially undesirable effects. Therefore, it is important that these models be used with care and caution, and be analyzed for biases and socially undesirable effects before use.

3. Additionally, due to unavailability of sufficient computing resources, we were unable to evaluate large language models such as BLOOM, even though it covers 22 African languages.

4. Finally, even though AfroNLU has diverse tasks at the word and sentence level, these tasks only cover few African languages. We therefore encourage the creation of more datasets for downstream NLU tasks in more (and more diverse) African languages. We believe broader benchmarks will continue to be important for future progress in African NLP.

## 9 Ethics Statement and Wider Impacts

SERENGETI aligns with Afrocentric NLP where the needs of African people is put into consideration when developing technology. We believe SERENGETI will not only be useful to speakers of the languages supported, but also researchers of African languages such as anthropologists and linguists. We discuss below some use cases for SERENGETI and offer a number of broad impacts.

1. SERENGETI aims to address the lack of access to technology in about 90% of the world's languages, which automatically discriminates against native speakers of those languages. More precisely, it does so by focusing on Africa. To the best of our knowledge, SERENGETI is the first massively multilingual PLM developed for African languages and language varieties. A model with knowledge of 517 African languages, is by far the largest to date for African NLP.

2. SERENGETI enables improved access of important information to the African community in Indigenous African languages. This is especially beneficial for people who may not be fluent in other languages. This will potentially connect more people globally.

3. SERENGETI affords opportunities for language preservation for many African languages. To the best of our knowledge, SERENGETI consists of languages that have not been used for any NLP task until now. We believe that it can help encourage continued use of these languages in several domains, as well as trigger future development of language technologies for many of these languages.

4. To mitigate discrimination and bias, we adopt a manual curation of our datasets. Native speakers of Afrikaans, Yorùbá, Igbo, Hausa, Luganda, Kinyarwanda, Chichewa, Shona, Somali, Swahili, Xhosa, Bemba, and Zulu also manually evaluated a subset of the data to ensure its quality. The data collected for this work is taken from various domains to further ensure a better representation of the language usage of native speakers.

5. Although LMs are useful for a wide range of applications, they can also be misused. SERENGETI is developed using publicly available datasets that may carry biases. Although we strive to perform analyses and diagnostic case studies to probe performance of our models, our investigations are by no means comprehensive nor guarantee absence of bias in the data. In particular, we do not have access

to native speakers of most of the languages covered. This hinders our ability to investigate samples from each (or at least the majority) of the languages.


## Acknowledgements

MAM gratefully acknowledges support from Canada Research Chairs (CRC), the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-2018-04267), the Social Sciences and Humanities Research Council of Canada (SSHRC; 435-2018-0576; 895-2020-1004; 895-2021-1008), Canadian Foundation for Innovation (CFI; 37771), Digital Research Alliance of Canada,[5] UBC ARC-Sockeye,[6] Advanced Micro Devices, Inc. (AMD), and Google. Any opinions, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of CRC, NSERC, SSHRC, CFI, the Alliance, AMD, Google, or UBC ARC-Sockeye.


## References


Julien Abadji, Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2021. Ungoliant: An optimized pipeline for the generation of a very large-scale multilingual web corpus. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-9) 2021. Limerick, 12 July 2021 (Online-Event), pages 1 – 9, Mannheim. Leibniz-Institut für Deutsche Sprache.

Ife Adebara and Muhammad Abdul-Mageed. 2022. Towards afrocentric NLP for African languages: Where we are and where we can go. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3814–3841, Dublin, Ireland. Association for Computational Linguistics.

Ife Adebara, Muhammad Abdul-Mageed, and Miikka Silfverberg. 2022a. Linguistically-motivated Yorùbá-English machine translation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5066–5075, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2022b. AfroLID: A neural language identification tool for African languages.

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022a. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

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 M. 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. 2022b. Masakhaner 2.0: Africa-centric transfer learning for named entity recognition.

Alham Fikri Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, Jey Han Lau, and Sebastian Ruder. 2022. One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249, Dublin, Ireland. Association for Computational Linguistics.

Jesujoba Alabi, Kwabena Amponsah-Kaakyire, David Adelani, and Cristina Espana-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.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022a. Adapting


[5]https://alliancecan.ca
[6]https://arc.ubc.ca/ubc-arc-sockeye

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, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022b. Multilingual language model adaptive fine-tuning: A study on African languages.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2018. Machine bias. *Nieman reports*, 72(3/4):37.

Jochen S. Arndt. 2015. *Missionaries, Africans and the Emergence of Xhosa and Zulu as Distinct Languages in South Africa, 1800-54*. Ph.D. thesis. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2022-11-02.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Israel Abebe Azime and Nebil Mohammed. 2021. An Amharic News Text classification Dataset.

Solon Barocas and Andrew D. Selbst. 2016. Big data's disparate impact. *California law review*, 104(3):671–732.

Emily M. Bender. 2011. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6.

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

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR.

Pei-Xuan Cai, Yao-Chung Fan, and Fang-Yie Leu. 2022. Compare encoder-decoder, encoder-only, and decoder-only architectures for text generation on low-resource datasets. In *Advances on Broad-Band Wireless Computing, Communication and Applications*, pages 216–225, Cham. Springer International Publishing.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Bharathi Raja Chakravarthi and Vigneshwaran Muralidaran. 2021. Findings of the shared task on hope speech detection for equality, diversity, and inclusion. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 61–72, Kyiv. Association for Computational Linguistics.

Zewen Chi, Shaohan Huang, Li Dong, Shuming Ma, Saksham Singhal, Payal Bajaj, Xia Song, and Furu Wei. 2021. Xlm-e: Cross-lingual language model pre-training via ELECTRA.

James Clackson. 2007. *The Indo-European language family*, Cambridge Textbooks in Linguistics, page 1–26. Cambridge University Press.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020a. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020b. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Bernard Comrie. 2017. *Languages of the World*, chapter 2. John Wiley & Sons, Ltd.

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.

Denis Creissels, Gerrit J Dimmendaal, Zygmunt Frajzyngier, and Christa König. 2008. Africa as a morphosyntactic area. *A linguistic geography of Africa*, 86150.

Brian Daigle. 2021. Data protection laws in Africa: A Pan-African survey and noted trends. *J. Int'l Com. & Econ.*, page 1.

Davis David. 2020a. Swahili : News classification dataset.

Davis David. 2020b. Swahili : News classification dataset. The news version contains both train and test sets.

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.

Mountaga Diallo, Chayma Fourati, and Hatem Haddad. 2021. Bambara language dataset for sentiment analysis.

Gerrit J. Dimmendaal. 2020. 364 Nilo-Saharan and Its Limits. In *The Oxford Handbook of African Languages*. Oxford University Press.

Gerrit J Dimmendaal, Colleen Ahland, Angelika Jakobi, and Constance Kutsch Lojenga. 2019. Linguistic features and typologies in languages commonly referred to as 'Nilo-Saharan'. *Cambridge Handbook of African Languages*, pages 326–381.

Sumanth Doddapaneni, Gowtham Ramesh, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2021. A primer on pretrained multilingual language models.

Bonaventure F. P. Dossou, Atnafu Lambebo Tonja, Oreen Yousuf, Salomey Osei, Abigail Oppong, Iyanuoluwa Shode, Oluwabusayo Olufunke Awoyomi, and Chris Chinenye Emezue. 2022. Afrolm: A self-active learning-based multilingual pretrained language model for 23 African languages.

Matthew S Dryer. 2013. Order of subject, object and verb. the world atlas of language structures online, ed. by matthew s. dryer and martin haspelmath. leipzig: Max planck institute for evolutionary anthropology. *Online: https://wals. info*.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. 2011. Fairness through awareness. Technical report, Cornell University Library, arXiv.org.

David M Eberhard, F Simons Gary, and Charles D Fennig (eds). 2021. Ethnologue: Languages of the world. *Twenty-fourth edition*, Dallas, Texas: SIL International.

Allyson Ettinger. 2020. What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Virginia Eubanks. 2018. *Automating inequality: how high-tech tools profile, police, and punish the poor*, first edition. St. Martin's Press, New York, NY.

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

Eduard Fosch-Villaronga and Adam Poulsen. 2022. *Diversity and Inclusion in Artificial Intelligence*, pages 109–134. T.M.C. Asser Press, The Hague.

Zygmunt Frajzyngier. 2018. Afroasiatic languages.

Ana Freire, Lorenzo Porcaro, and Emilia Gómez. 2021. Measuring diversity of artificial intelligence conferences. In *Proceedings of 2nd Workshop on Diversity in Artificial Intelligence (AIDBEI)*, volume 142 of *Proceedings of Machine Learning Research*, pages 39–50. PMLR.

Ludwig Gerhardt. 2020. 125 Reflections on the History of African Language Classification. In *The Oxford Handbook of African Languages*. Oxford University Press.

David Gil and Antoinette Schapper, editors. 2020. *Austronesian Undressed: How and why languages become isolating*. John Benjamins.

Jeff Good. 2020. 138139 Niger-Congo, with a Special Focus on Benue-Congo. In *The Oxford Handbook of African Languages*. Oxford University Press.

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. Larger-scale Transformers for multilingual masked language modeling. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 29–33, Online. Association for Computational Linguistics.

Tom Güldemann, editor. 2018. *The Languages and Linguistics of Africa*. De Gruyter Mouton.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. Towards a unified view of parameter-efficient transfer learning. In *International Conference on Learning Representations*.

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.

Sara Hooker. 2021. Moving beyond "algorithmic bias is a data problem". *Patterns*, 2(4):100241.

Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. "You Sound Just Like Your Father" Commercial Machine Translation Systems Include Stylistic Biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1686–1690, Online. Association for Computational Linguistics.

Larry M Hyman. 2003. African languages and phonological theory. *Glot International*, 7(6):153–163.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, et al. 2021. Masakhaner: Named entity recognition for African languages. *arXiv e-prints*, pages arXiv–2103.

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

Odunayo Jude Ogundepo, Akintunde Oladipo, Mofetoluwa Adeyemi, Kelechi Ogueji, and Jimmy Lin. 2022. AfriTeVA: Extending ?small data? pretraining approaches to sequence-to-sequence models. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 126–135, Hybrid. Association for Computational Linguistics.

Johanita Kirsten. 2018. *Afrikaans*, pages 13–30. Palgrave Macmillan UK, London.

Nikita Kitaev, Steven Cao, and Dan Klein. 2019. Multilingual constituency parsing with self-attention and pre-training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505, Florence, Italy. Association for Computational Linguistics.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences of the United States of America*, 117(14):pp. 7684–7689.

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 Suárez, 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. 2021. Quality at a glance: An audit of web-crawled multilingual datasets. *arXiv preprint arXiv:2103.12028*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pretraining for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

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*.

Alex Boniface Makulilo. 2012. Privacy and data protection in Africa: a state of the art. *International Data Privacy Law*, 2(3):163–178.

Nina Markl. 2022. Language variation and algorithmic bias: understanding algorithmic bias in British English automatic speech recognition. In *Proceedings of 2022 5th ACM Conference on Fairness, Accountability, and Transparency (FAccT 2022)*, pages 521–534. ACM Association for Computing Machinery.

Joshua L Martin and Kevin Tang. 2020. Understanding racial disparities in automatic speech recognition: The case of habitual" be". In *INTERSPEECH*, pages 626–630.

Yaron Matras. 2009. *Contact languages*, Cambridge Textbooks in Linguistics, page 275–307. Cambridge University Press.

Josh Meyer, Lindy Rauchenstein, Joshua D. Eisenberg, and Nicholas Howell. 2020. Artie bias corpus: An open dataset for detecting demographic bias in speech applications. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6462–6468, Marseille, France. European Language Resources Association.

Margaret Mitchell, Dylan Baker, Nyalleng Moorosi, Emily Denton, Ben Hutchinson, Alex Hanna, Timnit Gebru, and Jamie Morgenstern. 2020. Diversity and inclusion metrics in subset selection. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 117–123, New York, NY, USA. Association for Computing Machinery.

Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Sebastian Ruder, Ibrahim Said Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Salahudeen Abdullahi, Anuoluwapo Aremu, Alipio Jeorge, and Pavel Brazdil. 2022. NaijaSenti: A Nigerian Twitter Sentiment Corpus for Multilingual Sentiment Analysis.

Nico Nassenstein. 2019. Kinyarwanda and Kirundi: On Colonial Divisions, Discourses of National Belonging, and Language Boundaries. *Modern Africa: Politics, History and Society*, 7(1):11–40.

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.

Antoine Nzeyimana and Andre Niyongabo Rubungo. 2022. KinyaBERT: a morphology-aware Kinyarwanda language model. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5347–5363, Dublin, Ireland. Association for Computational Linguistics.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science (American Association for the Advancement of Science)*, 366(6464):447–453.

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 E Onyenwe, Mark Hepple, Uchechukwu Chinedu, and Ignatius Ezeani. 2018. A basic language resource kit implementation for the IgboNLP project. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 17(2).

Ikechukwu E. Onyenwe, Mark Hepple, Uchechukwu Chinedu, and Ignatius Ezeani. 2019. Toward an effective Igbo part-of-speech tagger. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 18(4).

Wuraola Fisayo Oyewusi, Olubayo Adekanmbi, and Olalekan Akinsande. 2020. Semantic enrichment of Nigerian Pidgin English for contextual sentiment classification. *arXiv preprint arXiv:2003.12450*.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

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

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Victor Porkhomovsky. 2020. 269Afro-Asiatic Overview. In *The Oxford Handbook of African Languages*. Oxford University Press.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Zaline Makini Roy-Campbell. 2006. The state of African languages and the global language politics: Empowering African languages in the era of globalization. In *Selected proceedings of the 36th annual conference on African linguistics*, pages 1–13. Cascadilla Proceedings Project Somerville, MA.

Sebastian Ruder. 2022. The State of Multilingual AI. http://ruder.io/state-of-multilingual-ai/.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Elizabeth-Jane Pavlick, Suzana Ili'c, Daniel Hesslow, Roman Castagn'e, Alexandra Sasha Luccioni, Franccois Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Rose Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurenccon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa Etxabe, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris C. Emezue, Christopher Klamm, Colin Leong, Daniel Alexander van Strien, David Ifeoluwa Adelani, Dragomir R. Radev, Eduardo G. Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady ElSahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jorg Frohberg, Josephine L. Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro von Werra, Leon Weber, Long Phan, Loubna Ben Allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, Mar'ia Grandury, Mario vSavsko, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad Ali Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto L'opez, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, S. Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal V. Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Févry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiang Tang, Zheng Xin Yong, Zhiqing Sun, Shaked Brody, Y Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre Franccois Lavall'ee, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aur'elie N'ev'eol, Charles Lovering, Daniel H Garrette, Deepak R. Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, S. Osher Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdenvek Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ananda Santa Rosa Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Olusola Ajibade, Bharat Kumar Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David M. Lansky, Davis David, Douwe

Kiela, Duong Anh Nguyen, Edward Tan, Emily Baylor, Ezinwanne Ozoani, Fatim T Mirza, Frankline Ononiwu, Habib Rezanejad, H.A. Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jan Passmore, Joshua Seltzer, Julio Bonis Sanz, Karen Fort, Lívia Macedo Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, M. K. K. Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nourhan Fahmy, Olanrewaju Modupe Samuel, Ran An, R. P. Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas L. Wang, Sourav Roy, Sylvain Viguier, Thanh-Cong Le, Tobi Oyebade, Trieu Nguyen Hai Le, Yoyo Yang, Zachary Kyle Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Kumar Singh, Benjamin Beilharz, Bo Wang, Caio Matheus Fonseca de Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel Le'on Perin'an, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully A. Burns, Helena U. Vrabec, Iman I.B. Bello, Isha Dash, Ji Soo Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthi Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, María Andrea Castillo, Marianna Nezhurina, Mario Sanger, Matthias Samwald, Michael Cullan, Michael Weinberg, M Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patricia Haller, R. Chandrasekhar, R. Eisenberg, Robert Martin, Rodrigo L. Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Pratap Bharati, T. A. Laud, Th'eo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yashasvi Bajaj, Y. Venkatraman, Yifan Xu, Ying Xu, Yun chao Xu, Zhee Xao Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2022. Bloom: A 176b-parameter open-access multilingual language model. *ArXiv*, abs/2211.05100.

Guillaume Segerer. 2008. Closed adjective classes and primary adjectives in African Languages. Working paper or preprint.

Iyanuoluwa Shode, David Ifeoluwa Adelani, and Anna Feldman. 2022. YOSM: A new Yorùbá Sentiment Corpus for Movie Reviews. *AfricaNLP 2022 @ICLR*.

Gabriele Sommer. 2020. 889 Pidgin and Creole Languages. In *The Oxford Handbook of African Languages*. Oxford University Press.

Xinying Song, Alex Salcianu, Yang Song, Dave Dopson, and Denny Zhou. 2021. Fast WordPiece tokenization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2089–2103, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ewan Sutherland. 2018. Digital privacy in Africa : cybersecurity, data protection & surveillance. *Data Protection & Surveillance (June 22, 2018)*.

Rachael Tatman. 2017. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain. Association for Computational Linguistics.

Rachael Tatman and Conner Kasten. 2017. Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions. In *Proc. Interspeech 2017*, pages 934–938.

S. Thara and Prabaharan Poornachandran. 2021. Transformer based language identification for Malayalam-English code-mixed text. *IEEE Access*, 9:118837–118850.

Andros Tjandra, Diptanu Gon Choudhury, Frank Zhang, Kritika Singh, Alexis Conneau, Alexei Baevski, Assaf Sela, Yatharth Saraf, and Michael Auli. 2021. Improved language identification through cross-lingual self-supervised learning.

UNESCO 66260. 2022. State of the art - indigenous languages in research webinar: concept note and agenda, 20 may 2022.

Asahi Ushio and Jose Camacho-Collados. 2021. T-NER: An all-round python library for Transformer-based named entity recognition. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 53–62, Online. Association for Computational Linguistics.

Rainer Vossen. 2020. 9091 African Language Types. In *The Oxford Handbook of African Languages*. Oxford University Press.

Alicia Beckford Wassink, Cady Gansen, and Isabel Bartholomew. 2022. Uneven success: automatic speech recognition and ethnicity-related dialects. *Speech Communication*, 140:50–70.

K. Williamson. 2006. Benue–Congo languages*. In Keith Brown, editor, *Encyclopedia of Language & Linguistics (Second Edition)*, second edition edition, pages 734–735. Elsevier, Oxford.

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.

Seid Muhie Yimam, Hizkiel Mitiku Alemayehu, Abinew Ayele, and Chris Biemann. 2020. Exploring Amharic sentiment analysis from social media texts: Building annotation tools and classification models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1048–1060, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained Transformer language models.

# Appendices

We provide an overview of the Appendix.
**Introduction**

- We share a large map of Africa showing the 517 Languages covered in our pretraining data in Figure A.1.
- We also share the scripts represented in our pretraining data in Table A.1.

**Literature Review**

- We provide a more extensive literature review in B. We discuss Afrocentric NLP, multilingualism in NLP, diversity and inclusion in NLP and multilingual language models.

**Pretraining Data** We discuss the pretraining data in more detain in Section C.
**Typology Information for AfroNLU** In Section D we discuss 6 families that cover the languages in 18 datasets in AfroNLU. For each family, we provide visualizations that cover any number of languages in the 18 datasets. We provide visualizations for:

- Afro-Asiatic in Figure D.1,
- Austronesian in Figure D.2,
- Creole in Figure D.3,
- Indo-European in Figure D.4,
- Niger-Congo in Figure D.5, and
- Nilo-Saharan in Figure D.6.

**Evaluation** We provide more information about the evaluations. We do the following:

- We show SERENGETI's performance on the Dev. set in Table E.1.
- We show SERENGETI's performance on each language in our multilingual datasets in Table E.4.
- We perform error analysis and show examples of errors in Table E.2. We also show confusion matrices for the news classification, sentiment analysis, and topic classification clusters in in Figure E.1, Figure E.2, and Figure E.3.
- We discuss the implications of imbalanced distribution and show confusion matrices for the news classification, sentiment analysis, and topic classification clusters in Figure E.4, Figure E.5, and Figure E.6.
- We show results from comparing SERENGETI with AfroLID and Franc on AfroLID test set in Table 7.
- Information about the languages in our pretraining data is provided in Table F.1, Table F.2 and Table F.3.
- We share statistics of the top ten languages with the largest data in SERENGETI and the ten languages with the least dataset in Table F.4.

**Genealogy /Language Contact Analysis** We further analyize our claim on the interaction of similar languages and zero-shot settings in Section F.

- We create a Figure highlighting the languages er perform analysis on in Figure E.7.
- We show the Jaccard similarity scores in Table 8.
- Next we show the results of each language in zero-shot settings and results for finetuning on BERT in Table 9.

## A Introduction

| Script | Languages |
|---|---|
| Ethiopic | Amharic, Basketo, Maale, *Oromo, Sebat Bet Gurage Tigrinya, Xamtanga |
| Arabic | Fulfude Adamawa, Fulfude Caka Tarifit |
| Vai | Vai |
| Coptic | Coptic |

Table A.1: Scripts represented in SERENGETI.

## B Literature Review

Representation learning is an integral part of modern NLP systems. It has significantly improved the state of the art in natural language understanding (NLU) and natural language generation (NLG). We now discuss Afrocentric NLP, Multilingualism in NLP, Diversity and Inclusion in NLP, MLMs, and LMs for African languages.

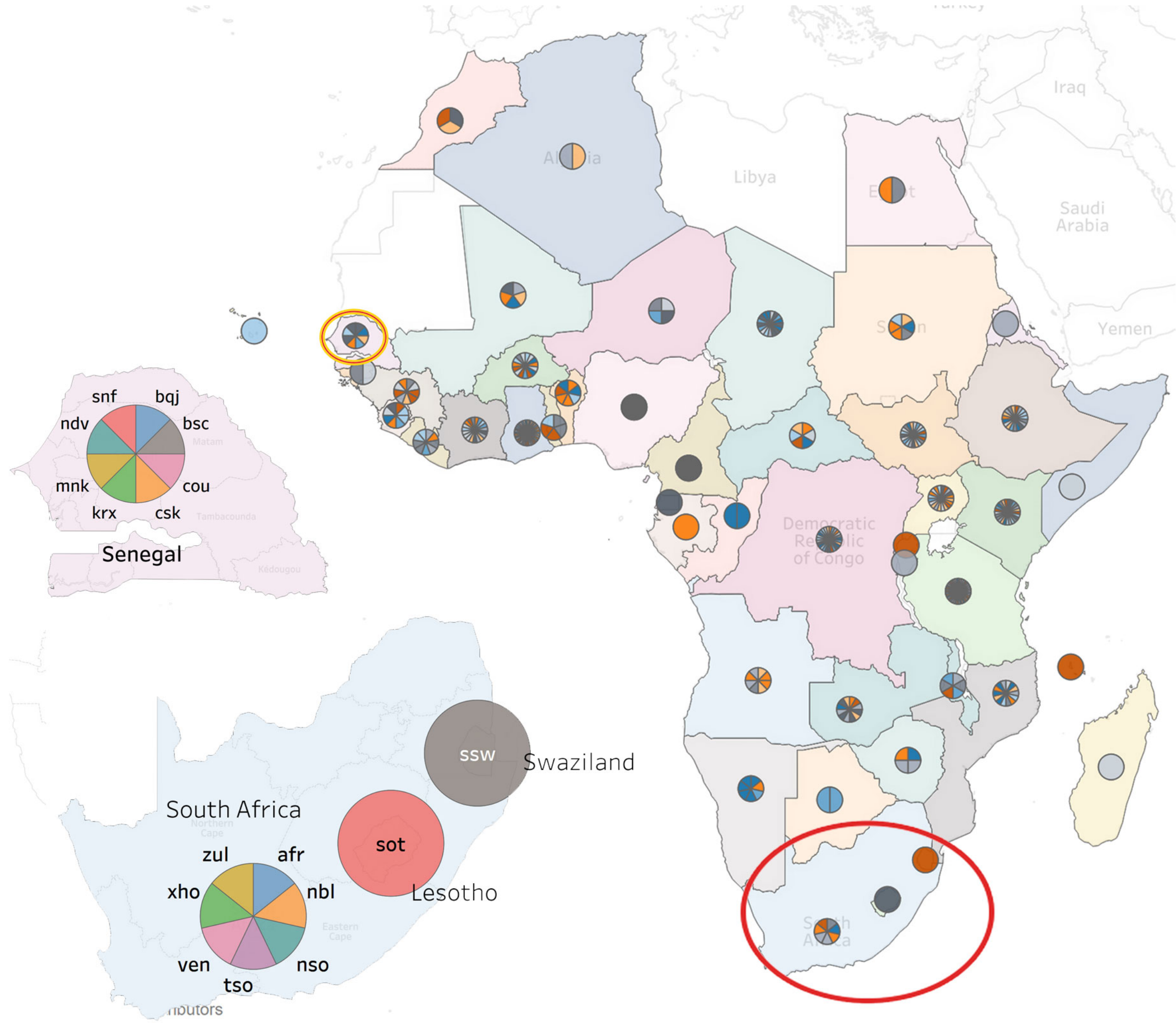



Figure A.1: All 517 languages in our dataset across the 50 African countries our data comes from. The language varieties are represented as colored pie shapes within each country. We zero in on South Africa, Lesotho, Swaziland, and Senegal to show detail.

### B.1 Afrocentric NLP

More than 2, 000 Indigenous languages are spoken in Africa, which is about a third of all languages spoken in the world (Eberhard et al., 2021). Unfortunately, the majority of these languages have not received any NLP attention to date. Rather, most NLP research has focused on higher resource languages. Most of these resourceful languages are typologically very different from Indigenous African languages. Methods used to develop technologies for these languages remain *Western-centric*, and may not be directly extensible to Indigenous African languages (Adebara and Abdul-Mageed, 2022). Existing NLP technologies also mostly function within the contexts of values and beliefs that reflect western societies and pose unique challenges if the technologies are applied within African communities.

Afrocentric NLP adopts a holistic approach to NLP throughout the life cycle of NLP policy making to model development and deployment. It discourages the current language data gas flaring policies that have led to the low resource status of many Indigenous African languages. Afrocentric NLP entails an understanding of the need for multidimensional policies that influence the language policy in education, media, government, and other domains to create ever-increasing, multi-domain, big data sources for NLP. During the archival and collection of language data, Afrocentric NLP necessitates respect of user consent, data sovereignty, wishes of local communities, and privacy (Sutherland, 2018; Daigle, 2021; Makulilo, 2012). For model development, approaches tailored to the unique typological features of African languages are of utmost priority. This also means development of models that understand simple to complex tones–a common feature in about 80% of African languages–serial verb constructions, and many other features (Hyman, 2003; Creissels et al., 2008). Afrocentric NLP also prioritizes deploying models in formats that people without programming experience can easily use. Furthermore, from an Afrocentric approach, development of certain NLP applications such as language models, language identification tools, spelling checkers, language specific keyboards, and machine translation systems is crucial to advance NLP for African languages.

### B.2 Multilingualism in NLP

Multilingualism, the ability to handle multiple languages within a single system or model, is becoming increasingly important as the amount of text and speech data in many languages increase. NLP systems capable of handling multiple languages can provide greater access to information and communication for people who speak languages other than those most commonly used or supported by NLP.

Multilingualism in NLP (Ruder, 2022) is mainly achieved through building (1) a single model trained on several languages (Devlin et al., 2019; Conneau et al., 2020) and (2) transfer learning (Raffel et al., 2020; He et al., 2022; Ruder et al., 2019). In the former, large transformer models have achieved state-of-the-art on many tasks while the latter has enabled the use of low-resource languages through finetuned on various NLP tasks. Due to lack of adequate (or good quality) pretraining data (Kreutzer et al., 2021), transfer learning is often the most accessible method for a few low resource languages. Unfortunately, about 94% of the world's languages are either *left-behinds*, in that it is probably impossible to build NLP resources for them, or *scraping-bys* with no labelled datasets (Joshi et al., 2020). For the left-behinds, labelled and unlabelled data is unavailable and even transfer learning approaches are beyond reach. So far, to the best of our knowledge, the largest multilingual model for African languages is pretrained on only 28 African languages (Dossou et al., 2022).

Most multilingual models are often trained with no more than 100 languages because increasing the number of language would mean decreasing its capacity to learn representations of each language (Conneau et al., 2020). Nevertheless, increasing model size was shown to ameliorate this problem (Goyal et al., 2021). In some cases, these benchmarks are translations from English (Artetxe et al., 2020; Nzeyimana and Niyongabo Rubungo, 2022; Ponti et al., 2020) and may not necessarily be a good evaluation for the languages. This is because translating from a source language may mask concept gaps and differences in linguistic constituents (Segerer, 2008) in the target language. That is, translations are at best approximations of the target language (Adebara and Abdul-Mageed, 2022; Joshi et al., 2020). For example, when translating into English (which marks (in)definiteness morphologically) from Yorùbá (which uses bare nouns but marks these features contextually), ambiguities arise (Adebara et al., 2022a).

For evaluation of multilingual models, several benchmarks have been created(Artetxe et al., 2020) with most of these supporting English and other high-resource languages. More recently, a few evaluation sets were introduced for African languages (Ifeoluwa Adelani et al., 2021; Shode et al., 2022; Niyongabo et al., 2020).We include these evaluation sets in our benchmark, which we henceforth refer to as AfroNLU.

When evaluating multilingual models, reporting model performance for each language in the benchmark is preferred because reporting the results as a single value on all languages may mask the model's performance on individual languages (Ruder, 2022). Large pre-training data, finetuning data, and evaluation benchmarks remain open challenging questions for achieving progress in multilingual NLP. For SERENGETI, we report results for each language in each benchmark across the 9 tasks we evaluate on.

### B.3 Diversity and Inclusion in NLP

Diversity relates to the level of variety within a system. It is the measure of distinctiveness between the various individuals within a group. Inclusion on the other hand relates to the level of representation or alignment of an individual within a group and the ability for that individual to function to its fullest ability (Fosch-Villaronga and Poulsen, 2022; Mitchell et al., 2020). Diversity and inclusion in NLP has gained increasing attention in recent years. In general, there is an acknowledgement that over-representation (and under-representation) of certain groups in the data used to train models (Mitchell et al., 2020) can be amplified by resulting technologies. This raises concerns about the technology and how it is that it can further existing biases and societal inequalities. But these biases can be exhibited in various ways beyond training data, including the algorithms implemented, the diversity of researchers and engineers developing the models, and the societal and cultural context in which they are used.

Although this is starting to change, often times most of the data exploited in NLP models come from closely related Western languages. Most of these languages are Indo-European (Aji et al., 2022; Joshi et al., 2020), and many of them share close geographic proximity and typology. In addition, the people who speak these languages have similar cultures. The implication is that several linguistic phenomena and typologies are under-represented in NLP data while those prevalent in Indo-European languages are over-represented (Chakravarthi and Muralidaran, 2021). About $88.38\%$ of the $2,679$ languages whose typology is described in WALS (Dryer, 2013) have not been used in NLP (Joshi et al., 2020). Many ideas and topics, alien to Western cultures have also never been seen (Adebara and Abdul-Mageed, 2022; Bender, 2011) in NLP data. African languages–and indeed many low resource languages–have rich linguistic typology, probably not seen in any other language in the world (Bender, 2011). An obvious problem with the current lack of diversity in NLP data is that the methods and models developed have overfit to these Indo-European typologies and cannot generalize to other typologies. Similarly, machine translation systems have been found to exhibit gender, racial (Bolukbasi et al., 2016; Caliskan et al., 2017; Chakravarthi and Muralidaran, 2021) and stylistic biases (Hovy et al., 2020) in their outputs perpetuated through the data used for training.

A number of studies have also found that algorithms could exhibit biases (Hooker, 2021; Buolamwini and Gebru, 2018; Dwork et al., 2011). For example, a recent study that investigated performance of Amazon Transcribe and Google Speech-To-Text on British English reported notably higher error rates for second language speakers of different varieties of British English (Markl, 2022). In another study, an evaluation of automatic speech recognition systems show substantial performance differences between 'standard' US English and African American English (AAE) varieties (Koenecke et al., 2020). In this study, commercial ASR systems developed by Amazon, Apple, Google, IBM, and Microsoft were evaluated and higher rates of errors were recorded for speakers of AAE than speakers of standard US varieties. Similar studies have also recorded higher errors in non-white users of English (Wassink et al., 2022; Martin and Tang, 2020). Other studies also reported differences in the performance of Youtube's automatic caption in different settings. One study reported higher accuracy in the transcriptions of US English compared with Indian English (Meyer et al., 2020). Another reported lower accuracy scores for women and speakers of Scottish English (Tatman, 2017) and non-white speakers of English (Tatman and Kasten, 2017).

Apart from data and algorithmic biases, the diversity crises in AI research is also argued to perpetuate historical biases (Freire et al., 2021). A more inclusive and diverse workforce could promote the exploration of questions and solutions beyond currently investigated research questions (Fosch-Villaronga and Poulsen, 2022). Several initiatives have been adopted to increase diversity in AI, including providing travel grants to marginalized communities to attend conferences, creating mentoring opportunities, special workshops, and community diversity chairs. A number of organizations have also been developed to promote diversity and inclusion in AI and NLP, such as Masakhane, Black in AI, LatinX in AI.

The impact of using biased systems in decision making have been extensively studied. Algorithmic decision-making using biased systems have been shown to have significant discriminatory effects in health (Obermeyer et al., 2019; Eubanks, 2018), employment (Barocas and Selbst, 2016), housing (Buolamwini and Gebru, 2018; Barocas and Selbst, 2016), government benefit allocation (Eubanks, 2018), policing (Buolamwini and Gebru, 2018; Barocas and Selbst, 2016; Angwin et al., 2018), and freedom (Angwin et al., 2018). Lack of diversity also has implication on access to technology. Currently, due to the use of a few high resource languages in NLP, there is limited global access to important applications such as machine translation, speech processing, information retrieval, and sentiment analysis. These technologies play an important role in ensuring a language thrives and offer major contributions to ongoing communication, literacy, education, and translation efforts in communities worldwide. These languages which have barely been used for NLP, usually referred to as low-resource languages, represent more than 90% of the world's $7,000$ languages (Joshi et al., 2020). The current focus of NLP on resource-rich languages does also have aggravating effects on the language endangerment problem which has been of serious concern for linguistics and language policy around the world. An alarming $50-90\%$ of languages have been envisaged to go extinct by the end of the century due to the domination by some of these resource-rich languages (Besacier et al., 2014).

Overall, diversity and inclusion in NLP remain active areas of research and comprise pressing issues of international significance. SERENGETI contributes to diversity and inclusion in NLP as follows: **(1)** We develop SERENGETI, a suite of massively, multilingual language models that support 517 African languages and language varieties. To the best of our knowledge, more than $400$ of these languages have never been represented in any language model to date. **(2)** The languages we support belong to 14 language families. **(3)** We provide a massive benchmark covering 28 languages across eight different tasks.

### B.4 Multilingual Language Models

MLMs have proven effective for cross-lingual NLU and NLG, often outperforming monolingual language models (Conneau et al., 2020). Different objectives have been adopted for training (Doddapaneni et al., 2021), using Transformer architectures. These LMs use one of the three different variants of Transformer architectures–encoder-decoder, encoder-only and decoder-only (Cai et al., 2022).

In the encoder-decoder models, input is encoded by the encoder side and the decoder conducts the operation to predict the sequence one token at a time or just reconstruct it by denoising. MBART (Liu et al., 2020), AfriTeva (Jude Ogundepo et al., 2022), M2M100 (Fan et al., 2020), and MT5 (Xue et al., 2021) are representatives for this architecture. Encoder-only models use only the encoder part of the transformer architecture, while decoder-only models use its decoder only. Some examples of encoder-only models are BERT (Devlin et al., 2019), XLMR (Conneau et al., 2020), and Electra (Chi et al., 2021), while BLOOM (Scao et al., 2022), GPT (Radford et al., 2018, 2019; Brown et al., 2020b), OPT (Zhang et al., 2022) are examples of decoder-only models. Most LMs developed for African languages use an encoder-only architecture, except AfriTEVA and AfroT5 which use encoder-decoder architectures.

These models are further finetuned on specific tasks. Finetuning has demonstrated its effectiveness on various NLU and NLG downstream tasks including part of speech tagging (Conneau et al., 2020), named entity recognition (Ushio and Camacho-Collados, 2021; Conneau et al., 2020), and question answering (Conneau et al., 2020). Finetuning follows a transfer learning approach which attempts to transfer knowledge from other sources to benefit a current task. This is based on the premise that previous knowledge may improve solutions for a current task (Pan and Yang, 2010; Raffel et al., 2020; He et al., 2022; Ruder et al., 2019). Transfer learning allows the domains, tasks, and distributions used in training and testing to be different thereby enabling a new task to leverage previously acquired domain knowledge. Potential benefits include faster learning, better generalization, and a more robust system. In the real world, we find many examples of transfer learning where humans transfer previous knowledge while learning or performing a task. For instance, knowing how to play the piano may facilitate learning to play the guitar and knowing how to ride a bicycle may facilitate learning to ride a motorbike. Finetuning is thus done by reusing the LM's parameters as a starting point, while adding one task-specific layer trained from scratch. Finetuning can be done on an individual or joint basis (Kitaev et al., 2019). In the former, a model is finetuned on single language for a specific downstream task. In the later, training data from a combination of multiple languages can be jointly finetuned in a single model.

## C Pretraining Data

We provide details of our pretraining data below:
**Religious Domain.** Our religious data is taken from online Bibles, Qurans, and data crawled from the Jehovah's witness website. We also include religious texts from the book of Mormon.
**News Domain.** We collect data from online newspapers (Adebara and Abdul-Mageed, 2022) and news sites such as Voice of America, Voice of Nigeria, BBC, Global voices, and DW news sites. We collect local newspapers from 27 languages from across Africa.
**Government Documents.** We collect government documents South African Centre for Digital Language Resources (SADiLaR), and the Universal Declaration of human rights (UDHR) in multiple languages.

**Health Documents.** We collect multiple health documents from the Department of Health, State Government of Victoria, Australia. We collect documents in Amharic, Dinka, Harari, Oromo, Somali, Swahili, and Tigrinya.

**Existing Corpora.** We collect corpora available on the web for different African languages, including from Project Gutenberg for Afrikaans, South African News data. for Sepedi and Setswana, OSCAR (Abadji et al., 2021) for Afrikaans, Amharic, Somali, Swahili, Oromo, Malagasy, and Yoruba. We also used Tatoeba for Afrikaans, Amharic, Bemba, Igbo, Kanuri, Kongo, Luganda, Malagasy, Sepedi, Ndebele, Kinyarwanda, Somali, Swahili, Tsonga, Xhosa, Yoruba, and Zulu; Swahili Language Modelling Data for Swahili; Ijdutse corpus for Hausa; Data4Good corpora for Luganda, CC-100 for Amharic, Fulah, Igbo, Yoruba, Hausa, Tswana, Lingala, Luganada, Afrikaans, Somali, Swahili, Swati, North Sotho, Oromo, Wolof, Xhosa, and Zulu; Afriberta-Corpus for Afaan / Oromo, Amharic, Gahuza, Hausa, Igbo, Pidgin, Somali, Swahili, Tigrinya and Yoruba; mC4 for Afrikaans, Amharic, Hausa, Igbo, Malagasy, Chichewa, Shona, Somali, Sepedi, Swahili, Xhosa, Yoruba and Zulu.

# D Typology Information for AfroNLU

SERENGETI consists of languages from 14 families including: Afro-Asiatic, Austronesean, Creole-English, Creole-French, Creole-Kongo, Creole-Ngbandi, Creole-Portuguese, khoe-kwadi-Hainum, khoe-kwadi-Nama khoe-kwadi-Southwest, Indo-European, Niger-Congo, and Nilo Saharan. We discuss the classes from AfroNLU which includes Afro-Asiatic, Austronesian, Creole-English, Niger-Congo, and Nilo-Saharan.

## D.1 Afro-Asiatic

Afro-Asiatic (*aka* Hamito-Semitic) is one of the language families of Africa. It consists of five or six branches: Berber, Chadic, Cushitic, Egyptian, Omotic (or a single Cush-Omotic), and Semitic(Porkhomovsky, 2020; Comrie, 2017). Many Afro-Asiatic languages are spoken in Central, East, North, and West Africa. They are also spoken in the Middle East and in scattered communities in Europe, the United States, and the Caucasus (Frajzyngier, 2018). In Figure D.1, we show relationship between the Afro-asiatic languages in AfroNLU.

## D.2 Austronesian

Austronesian languages are found along Mainland Southeast Asia, through Indonesia, Western New Guniea, and the Madagascar area in Africa (Eberhard et al., 2021). Many of them have been shown to exhibit an isolating word structure. This means that the words in these languages are of minimal morphological complexity (Gil and Schapper, 2020). In Figure D.2, we show the geneology for Malagasy, the only Austronesian language in our benchmark.

## D.3 Creole

A creole language is one spoken initially only in situations of contact between speakers of two or more mutually unintelligible languages, and not as a language within an ethnic group (Sommer, 2020). Historically, creoles have evolved along trade routes or in colonized communities particularly when several groups of people without a common lingua franca are forced to communicate in the presence of a dominant language. Creole languages therefore often include lexical items and grammatical features from multiple contact languages. Usually, one dominant language that is also referred to as the *lexifier* language contributes a majority of the vocabulary. Creole languages are classified based on their geographical location and are further grouped according to their main lexifier languages, their presumed origins, and the major languages with which they are in contact (i.e., *contact* languages). Figure D.3 shows the geneology for Nigerian Pidgin, the only Creole in our pretraining collection.

## D.4 Indo-European

Afrikaans is the only "Indigenous" Indo-European language spoken in Africa. Although it may also be viewed as not being truly Indigenous to Africa (Kirsten, 2018). Indo-European languages were originally domiciled in Europe, Iran, Turkey, Western Asia and India (Clackson, 2007; Eberhard et al., 2021; Comrie, 2017; Kirsten, 2018). However, due to migration, Indo-European languages are spoken around the world. In 2003, over 2.5 billion people spoke an Indo-European language (Clackson, 2007). In Figure D.4, we show the geneology for Afrikaans.

## D.5 Niger-Congo

Niger-Congo, also referred to as Niger-Kordofanian, is the largest language family

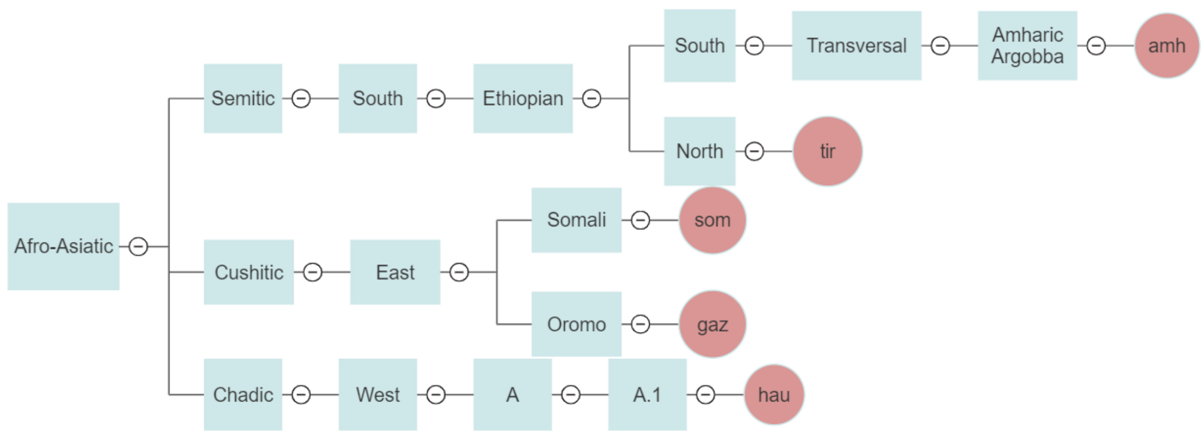



Figure D.1: Afro-Asiatic languages in SERENGETI pretraining data. Amharic (amh), Hausa (hau), Oromo (gaz), Somali (som) and Tigrinya (tir) are presented in red circles.

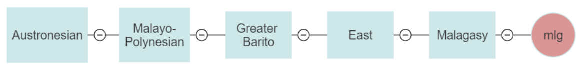



Figure D.2: Austroneasean language family consisting of Malagasy (mlg).


Figure D.3: SERENGETI pretraining data has one creole language, Nigerian Pidgin, indicated with ISO-639-3 code pcm.


Figure D.4: Indo-European language family consisting of Afrikaans (afr).

in Africa (Good, 2020; Comrie, 2017). It consists of the highest number of languages and speakers in Africa. Niger-Congo languages spread across sub-Saharan Africa, with Benue-Congo, including Bantu languages dominating the southern part of the continent. Figure D.5 shows the Niger-congo languages in our collection. Although we use similar colours for languages which are sisters of the same parent, only some of those languages are mutually intelligible. That is speakers of each individual language understand each other's language without learning it. Specifically, Kinyawanda (kin) and Kirundi (run) are mutually intelligible (Nassenstein, 2019). Ndebele, Siswati, Xhosa, and Zulu also share various levels of intelligibility mutually intelligible (Arndt, 2015; Roy-Campbell, 2006). Sepedi, Sotho, and Tswana also share some levels of mutual intelligibility (Roy-Campbell, 2006).

### D.6 Nilo-Saharan

Nilo-Saharan is subdivided into four branches that include North Eastern, Central Sudanic and two disputed branches–Songhay and Koman (Dimmendaal et al., 2019; Dimmendaal, 2020; Comrie, 2017). These branches are further divided into other subgroups, languages, and dialects. Nilo-Saharan languages are spoken predominantly by eastern and central African pastoralists, and includes in its main Chari-Nile branch the Central Sudanic and Eastern Sudanic (also called Nilotic) languages. Figure D.6 shows the Nilo-saharan languages in our pretraining data.

## E Evaluation

### E.1 Performance Analysis

In this section, we provide more information about our evaluation procedure and results using visualizations and tables. Figure E.1 shows the confusion matrix for the news classification cluster. Figure E.2 shows the performance of SERENGETI on the sentiment analysis cluster. Each confusion matrix represents each dataset in the sentiment analysis cluster. In Figure E.3, we show SERENGETI performance on each category in the topic classification datasets.

### E.2 Error Analysis

In the sentiment analysis cluster, best performance is recorded for positive categories while negative categories have the worst performance. A fine-grained analysis of the Yoruba sentiment dataset found that SERENGETI failed to correctly categor-

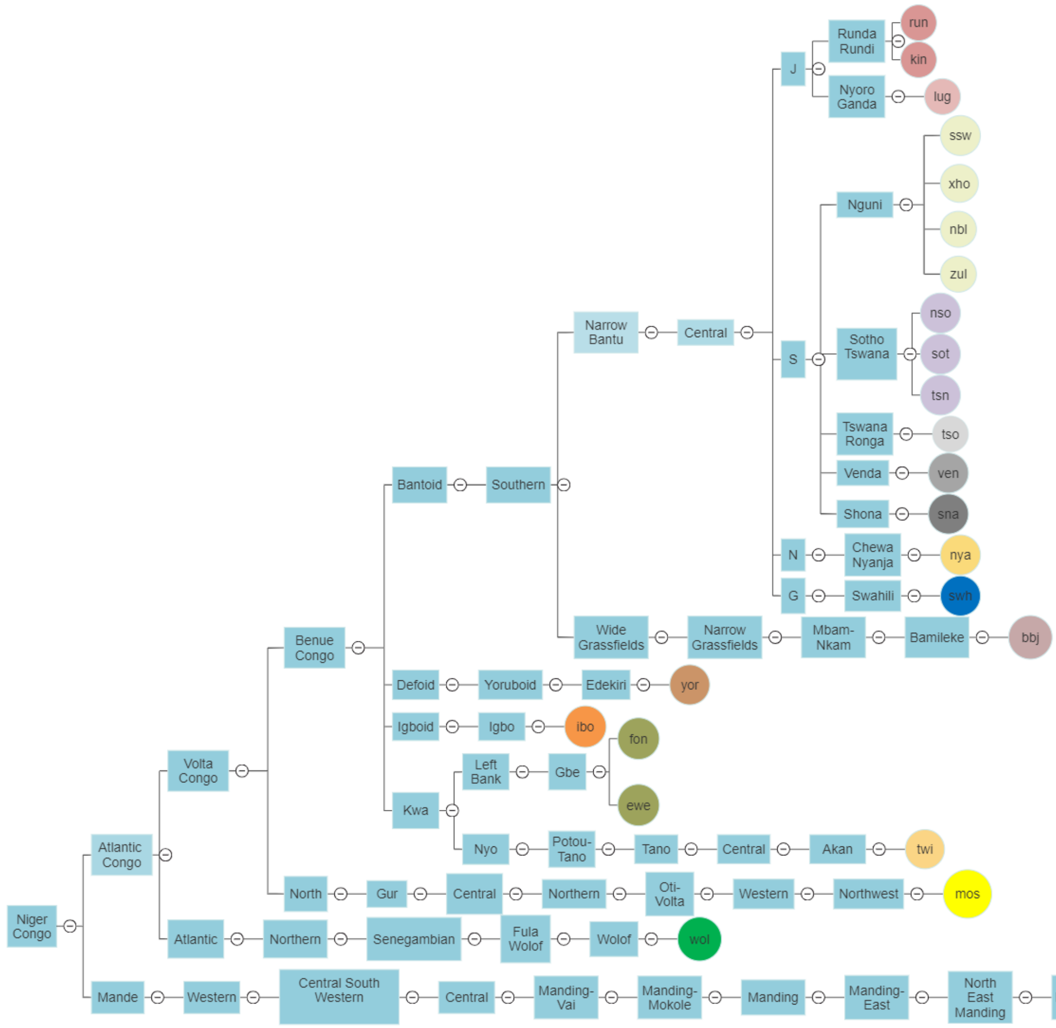



Figure D.5: Niger Congo Languages in AfroNLU benchmark. Languages which are siblings of the same parent are presented in similar colours.

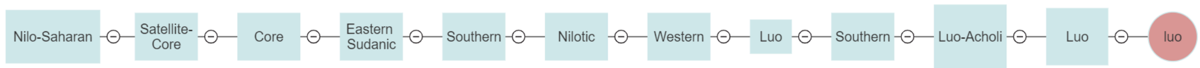



Figure D.6: Nilo Saharan language family with Luo (luo)

ize sentiment if the polarity item(s) were not seen in training, can be associated with both positive and negative sentiments, the polarity item(s) is a negation, or if ambivalent markers are present in the sentence. We provide a table showing examples of each type of error we found in Table E.2 in the Appendix. For the news classification task, politics and tourism are the best performing classes while education and relationships have the worst performance on kirnews and kinnews respectively. It is important to mention that the worst performing categories do not have the smallest data sizes. For the topic classification, the best performance is on the world class for Hausa topic modelling while entertainment and sport have best performance for Yoruba. The worst performance is on Nigeria and health for Hausa and Yoruba topic datasets respectively.

### E.3 Imbalanced Distribution

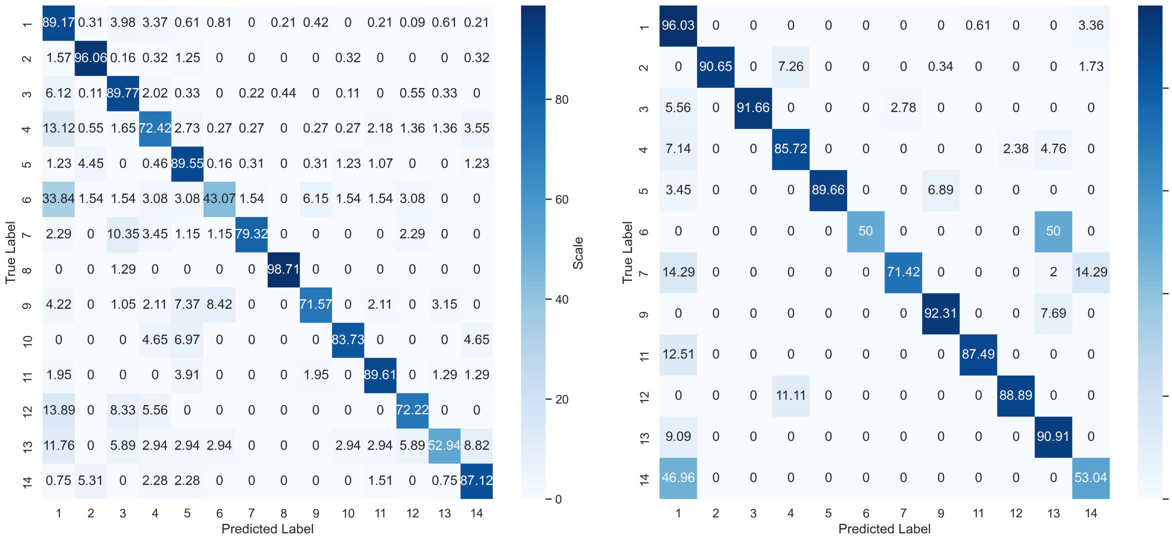
We find imbalances in the class distributions for all datasets except YOSM. We find a positive correlation between the size of each category in a dataset and the model accuracy. The larger the number of examples in a specific class, the better the accuracy,

| Cluster | Task | SOTA | XLMR | mBERT | Afro-XLMR | AfriBERTa | Serengeti-E110 | Serengeti-E250 | Serengeti |
|---|---|---|---|---|---|---|---|---|---|
| NER | masakaner-v1 | 84.8$^{\pm0.3}$ | 85.59 $^{\pm0.20}$ | 82.82 $^{\pm0.10}$ | 87.79 $^{\pm0.33}$ | 85.19 $^{\pm0.08}$ | 86.11$^{\pm0.27}$ | 86.42 $^{\pm0.26}$ | **88.82** $^{\pm0.18}$ |
| | masakaner-v2 | 85.7$^{\pm0.1\star}$ | 87.00 $^{\pm0.12}$ | 85.07$^{\pm0.83}$ | 87.46 $^{\pm0.06}$ | 86.19 $^{\pm0.11}$ | 86.51 $^{\pm0.22}$ | 86.81 $^{\pm0.24}$ | **88.98** $^{\pm0.20}$ |
| | masakaner-east | — | 83.52 $^{\pm1.03}$ | 82.85 $^{\pm0.42}$ | 87.28 $^{\pm0.68}$ | 83.33 $^{\pm0.56}$ | 85.64 $^{\pm0.50}$ | 87.12 $^{\pm0.62}$ | **88.09** $^{\pm0.57}$ |
| | masakaner-eastwest | — | 87.70 $^{\pm0.30}$ | 87.29 $^{\pm0.33}$ | 89.34 $^{\pm0.07}$ | 87.77 $^{\pm0.34}$ | 88.14 $^{\pm0.26}$ | 88.96$^{\pm0.15}$ | **90.38** $^{\pm0.17}$ |
| | masakaner-west | — | 89.77 $^{\pm0.53}$ | 90.28 $^{\pm0.46}$ | 89.97 $^{\pm0.23}$ | 89.36$^{\pm0.46}$ | 88.24 $^{\pm0.52}$ | 89.44 $^{\pm0.56}$ | **91.58** $^{\pm0.08}$ |
| | nchlt-ner | — | 72.19 $^{\pm0.13}$ | 71.44 $^{\pm0.07}$ | 73.22 $^{\pm0.2}$ | 69.25 $^{\pm0.25}$ | 65.67 $^{\pm0.07}$ | 65.86 $^{\pm0.16}$ | **73.81** $^{\pm0.18}$ |
| | yoruba-twi-ner | — | 57.40 $^{\pm2.51}$ | 75.35 $^{\pm0.78}$ | 68.02 $^{\pm2.01}$ | **82.40** $^{\pm0.04}$ | 65.6 $^{\pm2.87}$ | 62.45 $^{\pm1.04}$ | 79.68 $^{\pm1.42}$ |
| | wikiann | — | 84.82 $^{\pm0.24}$ | 84.68 $^{\pm0.85}$ | **87.00** $^{\pm1.12}$ | 84.58 $^{\pm0.46}$ | 84.21 $^{\pm0.12}$ | 85.64 $^{\pm0.36}$ | 86.91 $^{\pm0.31}$ |
| Phrase Chunking | phrase-chunk | — | 90.41 $^{\pm0.10}$ | 89.62 $^{\pm0.24}$ | 91.54 $^{\pm0.24}$ | 89.47 $^{\pm0.22}$ | 91.99 $^{\pm0.02}$ | 91.70 $^{\pm0.27}$ | **92.01** $^{\pm0.18}$ |
| POS | igbo-pos | — | 85.40 $^{\pm0.04}$ | 85.31 $^{\pm0.16}$ | 85.23 $^{\pm0.26}$ | 85.35 $^{\pm0.07}$ | 85.39 $^{\pm0.14}$ | **85.54** $^{\pm0.12}$ | 85.36 $^{\pm0.18}$ |
| News | amharic-news | — | 85.83 $^{\pm0.56}$ | 60.83 $^{\pm0.91}$ | 85.97 $^{\pm0.34}$ | **87.03** $^{\pm0.35}$ | 86.37 $^{\pm0.42}$ | 86.13 $^{\pm0.20}$ | 86.84 $^{\pm0.32}$ |
| | kinnews | — | 76.5 $^{\pm0.91}$ | 77.98 $^{\pm0.41}$ | 79.15 $^{\pm0.57}$ | 78.21 $^{\pm0.41}$ | **80.09** $^{\pm0.68}$ | 79.54 $^{\pm1.00}$ | 79.32 $^{\pm1.49}$ |
| | kirnews | — | 53.77 $^{\pm2.54}$ | 66.87 $^{\pm1.48}$ | 86.77 $^{\pm1.49}$ | **86.72** $^{\pm0.21}$ | 73.63 $^{\pm6.66}$ | 83.18 $^{\pm1.3}$ | 85.39 $^{\pm2.73}$ |
| | swahili-news-v0.2 | — | 88.43 $^{\pm0.31}$ | 85.28 $^{\pm0.21}$ | 88.89 $^{\pm0.58}$ | 88.76 $^{\pm0.82}$ | 88.09 $^{\pm1.02}$ | 86.97 $^{\pm1.31}$ | **89.29** $^{\pm0.74}$ |
| Sentiment Analysis | bambara-v2 | — | 46.22 $^{\pm1.94}$ | **65.00** $^{\pm2.00}$ | 62.81 $^{\pm1.35}$ | 60.19 $^{\pm1.61}$ | 60.50 $^{\pm0.94}$ | 63.90 $^{\pm3.5}$ | 63.17 $^{\pm0.51}$ |
| | pidgin-tweet | — | 69.99 $^{\pm0.41}$ | 69.00 $^{\pm0.44}$ | **71.41** $^{\pm0.16}$ | 69.47 $^{\pm0.84}$ | 69.98 $^{\pm0.35}$ | 69.64 $^{\pm0.23}$ | 68.27 $^{\pm1.11}$ |
| | yosm | — | 81.18 $^{\pm1.63}$ | 83.99 $^{\pm0.49}$ | 85.50 $^{\pm0.87}$ | **87.47** $^{\pm0.53}$ | 85.33 $^{\pm0.76}$ | 83.00 $^{\pm1.32}$ | 84.83 $^{\pm2.93}$ |
| Topic | hausa-topic | — | 84.75 $^{\pm1.88}$ | 83.48 $^{\pm1.52}$ | 87.83 $^{\pm0.53}$ | 88.41 $^{\pm0.49}$ | 87.50 $^{\pm0.11}$ | 88.21 $^{\pm0.61}$ | **89.52** $^{\pm1.11}$ |
| | yoruba-topic | — | 64.37 $^{\pm3.15}$ | 82.81 $^{\pm1.56}$ | **86.60** $^{\pm1.21}$ | 85.74 $^{\pm2.23}$ | 78.11 $^{\pm4.55}$ | 73.07 $^{\pm3.38}$ | 83.58 $^{\pm1.68}$ |
| **AfroNLU Score** | | | 77.77 | 79.54 | 82.96 | 80.92 | 80.03 | 80.43 | **83.04** |

Table E.1: Performance of models on seven AfroNLU benchmark DEV datasets. ($F_1$) score is the evaluation metric. In QA task, we train the models on English squad TRAIN and DEV datasets. We exclude the QA from AfroNLU DEV datasets. We use a dash (-) for tasks without a known SOTA.

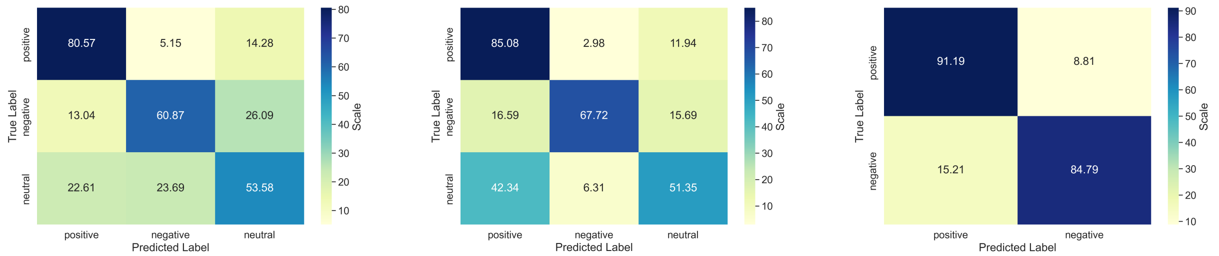
| Category | Sentence | Gold | Prediction |
|---|---|---|---|
| Ambivalence Markers | Kò burú ṣùgbọ̀n ó ti péjù | positive | negative |
| | Sinimá tì a lè pè nì ìràwọ́ sinimá tì ò ǹ<br>kọ mọ́nà mọ́nà ṣùgbọ̀n n tì kò nì ohun ámúyẹ ni. | negative | positive |
| Negation | Eré síse naa ko dára to, ìtàn naa kò yeni,<br>ní èrò tèmi òṣèré tó daa jù ni ìyá náà | negative | positive |
| | Ṣe oun tó o fẹ́. | negative | positive |
| Not seen in training | Wọn rí sinima yìí ṣe, àgbọ́dọ̀ wò ni | positive | negative |
| | Irú yádi fíímù. Mo kórìrá gbogbo dídágbé mi<br>nìkan kejì tì o. Ìdọtí ńlá! | negative | positive |
| Polarity item can be either positive or negative | Ìkìlọ̀. O ní láti wo ìparí eré yìí nítorí wípé ńkan<br>ṣẹlẹ̀ ní ìparí eré náà. | positive | negative |
| | Nìkan ní ìpò àwàdà Nollywood gbòòrò. Ṣé ó ní<br>ìdánílójú nítòótọ́. | negative | positive |

Table E.2: Error analysis of Yoruba Sentiment analysis dataset. The polarity items are highlighted in red.



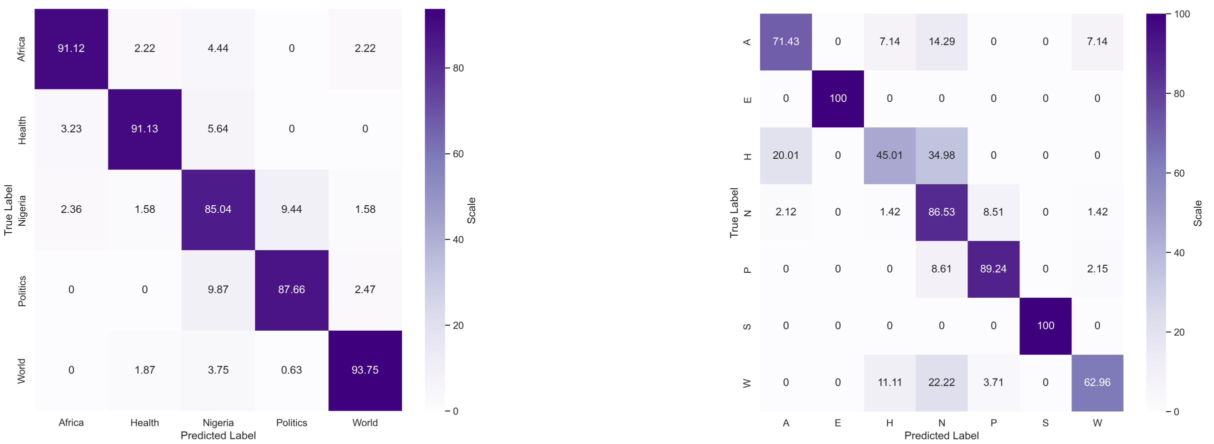

a) Kinnews b) Kirnews

Figure E.1: Confusion matrices showing the performance of SERENGETI for each categories in Kirnews and Kinnews classification datasets. The categories are (1) politics, (2) sports, (3) economy, (4) health, (5) entertainment, (6) history, (7) technology, (8) tourism, (9) culture, (10) fashion (11) religion, (12) environment, (13) education, and (14) relationship. Kirnews does not have Class 8 and 10.



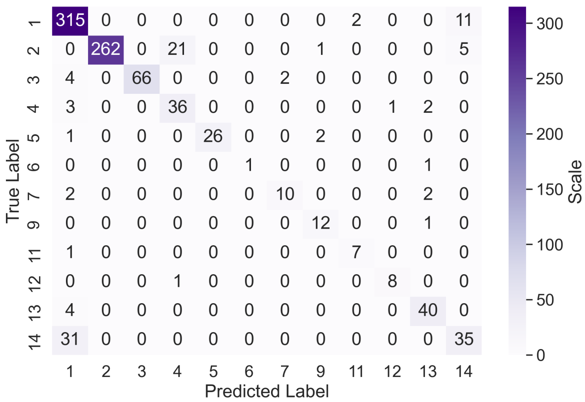

a) Bambara Sentiment Analysis b) Pidgin Tweets c) YOSM Sentiment

Figure E.2: Confusion matrices showing the performance of SERENGETI for each category in Bambara, Pidgin tweets, and YOSM datasets.

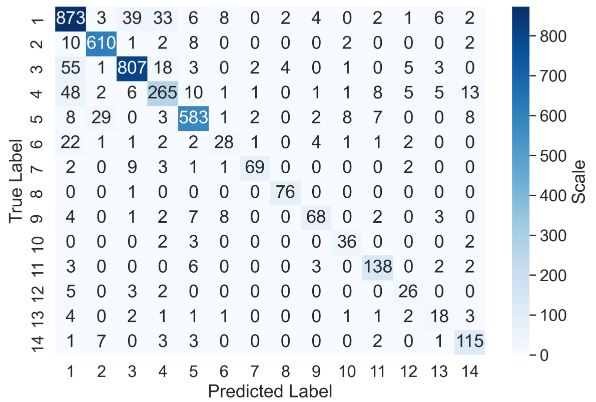



a) Hausa Topic Classification. b) Yoruba Topic Classification.

Figure E.3: Confusion matrices showing the performance of SERENGETI for each categories in Hausa and Yoruba topic classification datasets. A="Africa", E="Entertainment", H="Health", N="Nigeria", P="Politics", S="Sport", W="World" although we find a few exceptions. We provide confusion matrices that represents the sizes of each category and the performance of SERENGETI in Figures E.4, E.5, and E.6.

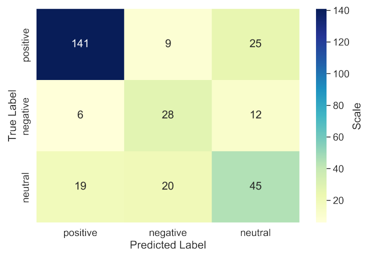



a) Kinnews

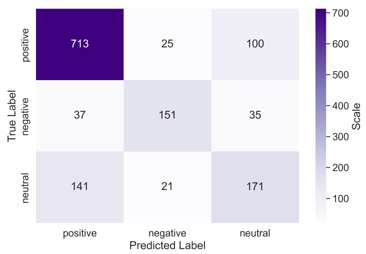



b) Kirnews

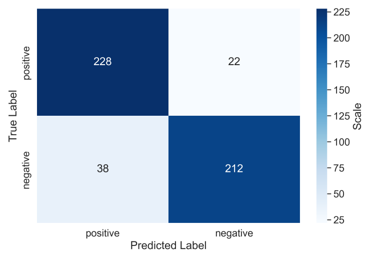
Figure E.4: Confusion matrices showing the performance of SERENGETI for each categories in Kirnews and Kinnews classification datasets.




a) Bambara Sentiment Analysis

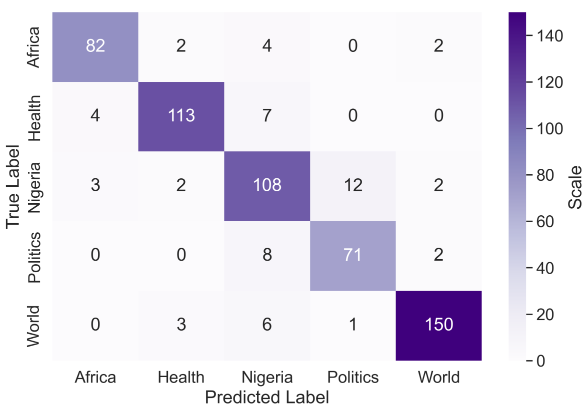



b) Pidgin Tweets

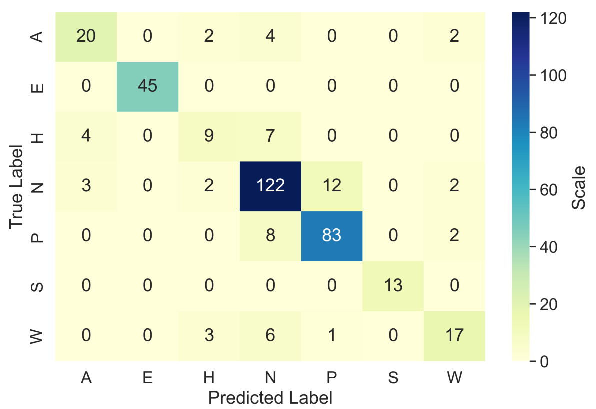



c) YOSM Sentiment

Figure E.5: Confusion matrices showing the performance of SERENGETI for each category in Bambara, Pidgin tweets, and YOSM datasets.




a) Hausa Topic Classification.




b) Yoruba Topic Classification.

Figure E.6: Confusion matrices showing the performance of SERENGETI for each categories in Hausa and Yoruba topic classification datasets. A="Africa", E="Entertainment", H="Health", N="Nigeria", P="Politics", S="Sport", W="World"

| ISO-639-3 | SERENGETI | AfroLID | Franc | ISO-639-3 | SERENGETI | AfroLID | Franc | ISO-639-3 | SERENGETI | AfroLID | Franc |
|---|---|---|---|---|---|---|---|---|---|---|---|
| aar | **100.00** | 96.00 | 74.00 | kde | **99.00** | 95.00 | 60.00 | pov | **98.00** | 93.00 | 82.00 |
| ada | **100.00** | **100.00** | 98.00 | kdh | **100.00** | 99.00 | 95.00 | run | **97.00** | 91.00 | 68.00 |
| afr | **100.00** | 97.00 | 81.00 | kea | **98.00** | 96.07 | 0.00 | sag | **100.00** | **100.00** | 30.00 |
| amh | **99.00** | 97.00 | 36.00 | kin | **94.00** | 89.00 | 47.00 | shk | **100.00** | **100.00** | 93.00 |
| bam | **92.00** | 70.00 | 30.00 | kmb | **98.00** | 94.00 | 71.00 | sna | **98.00** | 97.00 | 91.00 |
| bba | **100.00** | **100.00** | 83.00 | kng | **99.00** | 98.00 | 58.00 | som | **98.00** | 95.00 | 89.00 |
| bci | 97.00 | **98.00** | 92.00 | koo | **96.00** | **96.00** | **96.00** | sot | 92.00 | 88.00 | **93.00** |
| bem | **98.00** | 94.00 | 90.00 | kqn | **99.00** | 98.00 | 84.00 | ssw | **92.00** | 86.00 | 68.00 |
| bfa | **100.00** | 99.00 | 91.00 | kqs | **99.00** | 95.00 | 73.00 | suk | **100.00** | 99.00 | 34.00 |
| bin | **100.00** | 99.00 | 97.00 | ktu | **98.00** | 93.00 | 19.00 | sus | **99.00** | **99.00** | 96.00 |
| bum | **98.00** | 97.00 | 72.00 | lia | 98.00 | 97.00 | **100.00** | swh | **95.00** | 77.00 | 70.00 |
| cjk | **98.00** | 96.00 | 56.00 | lin | 98.00 | **99.00** | 98.00 | tem | **99.00** | **99.00** | 88.00 |
| crs | **97.00** | 96.00 | 83.00 | lot | **100.00** | 99.00 | 93.00 | tir | **100.00** | 99.00 | 97.00 |
| dag | **100.00** | **100.00** | **100.00** | loz | **100.00** | 95.00 | 92.00 | tiv | **100.00** | **100.00** | 99.00 |
| dga | 98.00 | **100.00** | 78.00 | lua | 98.00 | **99.00** | 87.00 | toi | **98.00** | **98.00** | 80.00 |
| dip | **98.00** | 93.00 | 86.00 | lue | **98.00** | 95.00 | 68.00 | tsn | **81.00** | 76.00 | 33.00 |
| dyu | 95.00 | **96.00** | 0.00 | lug | **96.00** | 87.00 | 64.00 | tso | 97.00 | **99.00** | 94.00 |
| ewe | 93.00 | **97.00** | **97.00** | lun | **97.00** | **97.00** | 86.00 | twi | **100.00** | **100.00** | 87.00 |
| fat | **98.00** | **98.00** | 94.00 | men | 98.00 | 98.00 | **99.00** | umb | **100.00** | 99.00 | 76.00 |
| fon | **98.00** | 97.00 | 92.00 | mfq | 92.00 | **95.00** | 88.00 | vai | **100.00** | **100.00** | **100.00** |
| fuf | **96.00** | 93.00 | 52.00 | mos | **99.00** | 97.00 | 90.00 | ven | **98.00** | 95.00 | 85.00 |
| fuv | **95.00** | 94.00 | 61.00 | nba | **100.00** | 99.00 | 61.00 | vmw | **98.00** | 97.00 | 95.00 |
| gaa | **98.00** | 95.00 | 97.00 | nbl | **79.00** | 74.00 | 47.00 | wol | **87.00** | 81.00 | 21.00 |
| gaz | 94.00 | 94.00 | **96.00** | ndo | **97.00** | 96.00 | 76.00 | xho | **75.00** | 67.00 | 30.00 |
| gjn | **100.00** | 98.00 | 99.00 | nso | **89.00** | 83.00 | 59.00 | xsm | **99.00** | **99.00** | 53.00 |
| gkp | 68.00 | 63.00 | **69.00** | nya | **99.00** | 92.00 | 75.00 | yor | **99.00** | 98.00 | 66.00 |
| hau | **95.00** | 88.00 | 77.00 | nym | 98.00 | **99.00** | 54.00 | zdj | **98.00** | 96.00 | 63.00 |
| ibb | **99.00** | 98.00 | 84.00 | nyn | **95.00** | 92.00 | 92.00 | zul | **68.00** | 50.00 | 40.00 |
| ibo | **97.00** | **97.00** | 88.00 | nzi | **100.00** | 97.00 | 98.00 | | | | |
| kbp | **100.00** | **100.00** | 98.00 | pcm | **96.00** | **96.00** | 82.00 | | | | |
| **SERENGETI Average f1_score:** 96.29 | | | | AfroLID Average f1_score: 91.63 | | | | Franc Average: f1_score 74.81 | | | |

Table E.3: $F_1$-scores for SERENGETI, AfroLID, and Franc on AfroLID's dataset for $88$ languages.

## F Detailed Geneaology and Language Contact Analysis

In this Section, we use Figures and Tables to provide evidence for the influence of similar languages in zero-shot settings. First, we highlight in purple the similar languages that we perform genealogy analysis on in Figure E.7. In the figure, the languages with mutual intelligibility are presented in similar coloured circles. To determine the significance of language similarity and language contact in our own zero-shot settings, we measure the Jaccard similarity between the pretraining data for the South African languages in AfroNLU (see Table 8). To calculate the Jaccard similarities, we removed digits, emojis, and punctuation marks. We do this to ensure that we reduce interference with the similarity scores. We find strong similarities between some of these languages as in the bolded examples in Table 8.

We find that although XLM-R, mBERT, and AfriBERTa are not trained on most most of these languages, we record high scores in zero-shot settings see Table E.4). We argue that XLM-R in addition to cross-lingual transfers from other languages acquires representation from afr and xho where xho alone shares more than 0.4 similarity with afr, nbl, nso, and zul. mBERT also learns representation from afr while AfriBERTa learns representations from Gahuza which is a code-mixed variety of kin and run. SERENGETI however, outperforms other models on these datasets indicating that learning the representation of each language improves performance.

Next, we finetune a BERT model and compare the performance of BERT with MBERT. We do this because BERT is a monolingual model and does not include any similar language in its representation. In Table 9, BERT significantly performs lower than MBERT in all languages in NCHLT-NER. BERT also has lower performance on the phrase-chunk dataset in all languages except on ssw, and ven.

This analysis is far from being conclusive and future work can further probe the influence of similar languages in more detail. This is necessary to evaluate to what extent similar languages have an influence on performance in zero-shot settings and why in zero shot settings, some monolingual models outperform multilingual ones. For example, in the case of ssw and ven.

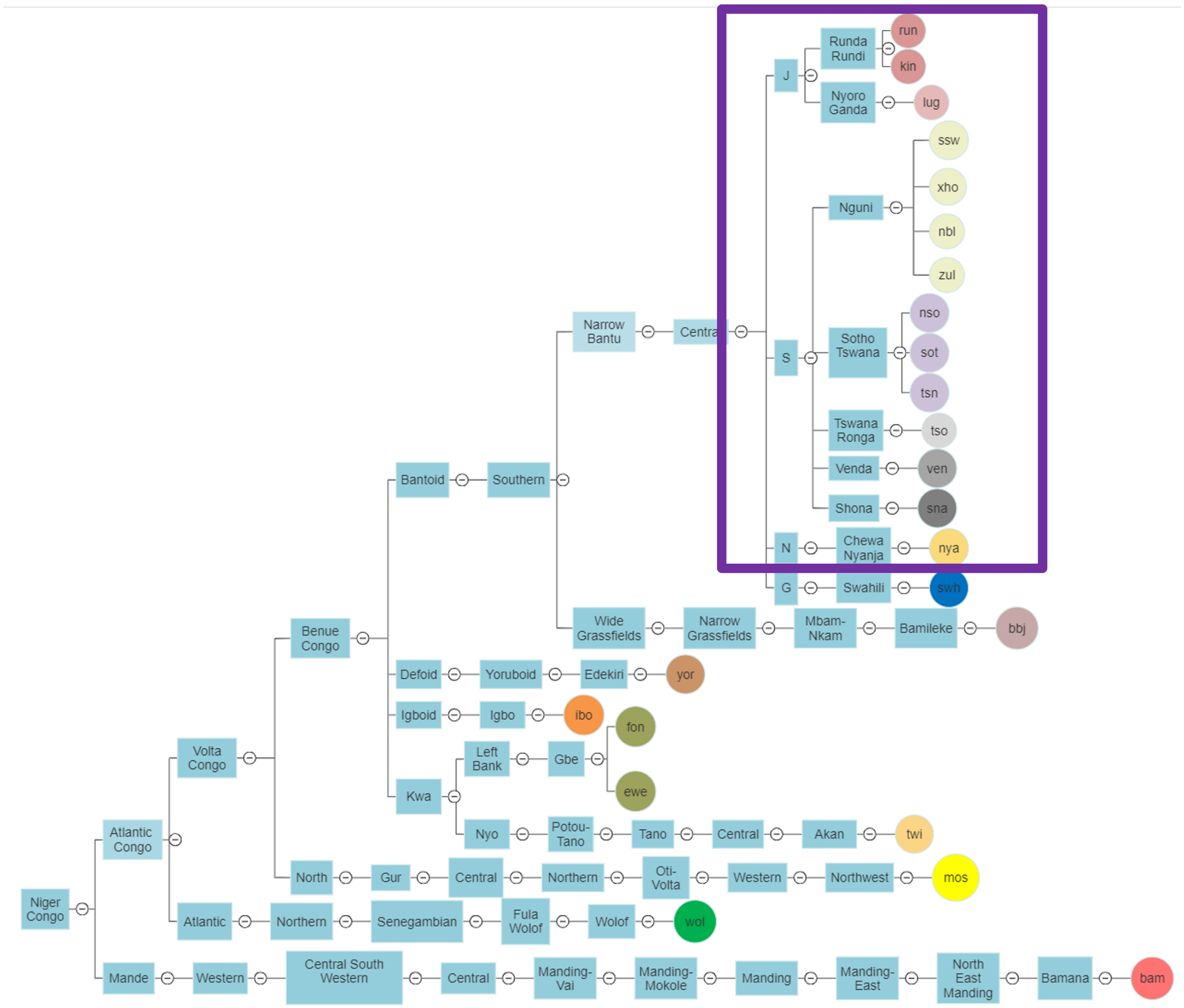



Figure E.7: A genetic classification of Niger-Congo languages in AfroNLU. We highlight in purple the list of languages relevant to our geneaology and language contact analysis. Languages which share stronger mutual intelligibility is represented in similar colours.

| Cluster | Dataset | Lang. | XLMR | mBERT | Afro-XLMR | AfriBERTa | SERENGETI |
|---|---|---|---|---|---|---|---|
| Named Entity Recognition (NER) | masakaner-v1 | amh | 73.98 $^{\pm0.64}$ | 0.0 $^{\pm0.0}$ | **77.38** $^{\pm0.47}$ | 69.61 $^{\pm0.76}$ | 74.26 $^{\pm0.54}$ |
| | | hau | 91.39 $^{\pm0.24}$ | 88.25 $^{\pm0.42}$ | 91.92 $^{\pm0.86}$ | 91.12 $^{\pm0.37}$ | **92.03** $^{\pm0.59}$ |
| | | ibo | 84.55 $^{\pm0.15}$ | 84.44 $^{\pm0.97}$ | 87.51 $^{\pm0.92}$ | 87.95 $^{\pm0.54}$ | **87.82** $^{\pm0.63}$ |
| | | kin | 73.54 $^{\pm0.35}$ | 71.02 $^{\pm1.34}$ | 78.46 $^{\pm0.34}$ | 75.07 $^{\pm0.51}$ | **78.56** $^{\pm0.34}$ |
| | | lug | 78.65 $^{\pm1.25}$ | 79.07 $^{\pm2.01}$ | 82.11 $^{\pm0.99}$ | 77.84 $^{\pm0.4}$ | **84.61** $^{\pm0.4}$ |
| | | luo | 74.28 $^{\pm1.87}$ | 74.47 $^{\pm0.08}$ | 75.20 $^{\pm1.23}$ | 70.76 $^{\pm1.57}$ | **77.28** $^{\pm1.61}$ |
| | | pcm | 88.89 $^{\pm0.56}$ | 88.88 $^{\pm0.91}$ | **90.07** $^{\pm0.18}$ | 87.65 $^{\pm0.43}$ | 89.65 $^{\pm0.63}$ |
| | | swa | 87.68 $^{\pm0.98}$ | 86.12 $^{\pm0.5}$ | 87.77 $^{\pm0.1}$ | 87.72 $^{\pm0.13}$ | **88.08** $^{\pm0.13}$ |
| | | wol | 63.4 $^{\pm0.68}$ | 64.25 $^{\pm1.66}$ | **68.09** $^{\pm1.65}$ | 60.9 $^{\pm1.69}$ | 66.26 $^{\pm1.47}$ |
| | | yor | 78.97 $^{\pm0.93}$ | 79.45 $^{\pm0.36}$ | **83.76** $^{\pm0.34}$ | 79.89 $^{\pm0.89}$ | 83.08 $^{\pm1.18}$ |
| | masakaner-v2 | bam | 80.66 $^{\pm0.99}$ | 79.2 $^{\pm1.43}$ | 81.04 $^{\pm0.31}$ | 78.55 $^{\pm0.42}$ | **82.11** $^{\pm0.53}$ |
| | | bbj | 72.82 $^{\pm1.07}$ | 62.44 $^{\pm0.59}$ | 73.31 $^{\pm0.74}$ | 71.97 $^{\pm1.61}$ | **73.66** $^{\pm0.87}$ |
| | | ewe | 88.54 $^{\pm0.23}$ | 84.19 $^{\pm1.12}$ | 89.58 $^{\pm0.54}$ | 86.97 $^{\pm0.4}$ | **89.75** $^{\pm0.14}$ |
| | | fon | 82.34 $^{\pm0.09}$ | 77.87 $^{\pm0.47}$ | 82.62 $^{\pm0.73}$ | 78.66 $^{\pm0.39}$ | **82.86** $^{\pm0.53}$ |
| | | hau | 86.09 $^{\pm0.61}$ | 82.66 $^{\pm1.46}$ | 87.29 $^{\pm0.67}$ | 86.14 $^{\pm0.38}$ | **87.33** $^{\pm0.62}$ |
| | | ibo | 89.67 $^{\pm0.28}$ | 84.04 $^{\pm1.09}$ | 91.99 $^{\pm0.11}$ | 91.56 $^{\pm0.36}$ | **92.28** $^{\pm0.21}$ |
| | | kin | 84.04 $^{\pm0.48}$ | 83.53 $^{\pm0.81}$ | **86.51** $^{\pm0.3}$ | 83.22 $^{\pm0.25}$ | 86.38 $^{\pm0.35}$ |
| | | lug | 86.18 $^{\pm0.22}$ | 85.78 $^{\pm1.41}$ | 88.17 $^{\pm0.56}$ | 85.32 $^{\pm0.49}$ | **89.24** $^{\pm0.37}$ |
| | | mos | 74.55 $^{\pm0.65}$ | 67.75 $^{\pm1.84}$ | **75.25** $^{\pm0.71}$ | 69.95 $^{\pm0.89}$ | 73.74 $^{\pm1.62}$ |
| | | nya | 90.23 $^{\pm0.14}$ | 88.6 $^{\pm0.65}$ | **91.84** $^{\pm0.23}$ | 88.83 $^{\pm0.11}$ | 91.29 $^{\pm0.19}$ |
| | | pcm | 89.11 $^{\pm0.1}$ | 87.90 $^{\pm1.0}$ | **89.27** $^{\pm0.4}$ | 87.81 $^{\pm0.45}$ | 88.77 $^{\pm0.37}$ |
| | | sna | 94.15 $^{\pm0.19}$ | 93.06 $^{\pm0.75}$ | 95.35 $^{\pm0.16}$ | 93.51 $^{\pm0.32}$ | **95.92** $^{\pm0.2}$ |
| | | swa | 92.37 $^{\pm0.05}$ | 91.09 $^{\pm0.33}$ | **93.06** $^{\pm0.14}$ | 92.43 $^{\pm0.11}$ | 92.87 $^{\pm0.33}$ |
| | | tsn | 85.69 $^{\pm0.89}$ | 85.02 $^{\pm0.85}$ | 88.24 $^{\pm0.26}$ | 83.58 $^{\pm0.79}$ | **88.43** $^{\pm0.1}$ |
| | | twi | 79.60 $^{\pm1.45}$ | 78.05 $^{\pm2.3}$ | 79.94 $^{\pm1.6}$ | 75.35 $^{\pm0.81}$ | **80.25** $^{\pm1.1}$ |
| | | wol | 85.14 $^{\pm0.34}$ | 83.65 $^{\pm1.11}$ | 84.60 $^{\pm0.4}$ | 81.68 $^{\pm0.38}$ | **85.97** $^{\pm0.43}$ |
| | | xho | 87.6 $^{\pm0.15}$ | 86.24 $^{\pm1.2}$ | **89.59** $^{\pm0.37}$ | 86.18 $^{\pm0.17}$ | 88.76 $^{\pm0.76}$ |
| | | yor | 86.56 $^{\pm0.36}$ | 83.45 $^{\pm1.63}$ | **88.91** $^{\pm0.27}$ | 87.45 $^{\pm0.17}$ | 87.99 $^{\pm0.61}$ |
| | | zul | 86.32 $^{\pm0.6}$ | 84.16 $^{\pm1.75}$ | 89.75 $^{\pm0.16}$ | 84.9 $^{\pm0.27}$ | **90.41** $^{\pm0.24}$ |
| | nchlt-ner | afr | 80.68 $^{\pm0.75}$ | 80.08 $^{\pm0.29}$ | 80.55 $^{\pm0.11}$ | 74.5 $^{\pm0.64}$ | **81.57** $^{\pm0.59}$ |
| | | nbl | 74.64 $^{\pm0.66}$ | 73.48 $^{\pm0.18}$ | 75.26 $^{\pm0.28}$ | 72.28 $^{\pm0.67}$ | **77.13** $^{\pm0.67}$ |
| | | nso | 77.0 $^{\pm1.23}$ | 78.75 $^{\pm0.45}$ | 80.13 $^{\pm0.51}$ | 75.45 $^{\pm1.09}$ | **80.69** $^{\pm0.64}$ |
| | | sot | 54.71 $^{\pm1.51}$ | 54.68 $^{\pm0.49}$ | 55.57 $^{\pm0.2}$ | 54.09 $^{\pm0.98}$ | **56.26** $^{\pm1.52}$ |
| | | ssw | 71.75 $^{\pm0.65}$ | 71.24 $^{\pm0.75}$ | 72.35 $^{\pm1.02}$ | 69.38 $^{\pm0.58}$ | **73.37** $^{\pm0.82}$ |
| | | tsn | 77.02 $^{\pm0.22}$ | 76.35 $^{\pm0.47}$ | 77.68 $^{\pm0.96}$ | 73.89 $^{\pm1.41}$ | **79.05** $^{\pm0.75}$ |
| | | tso | 74.24 $^{\pm0.08}$ | 72.95 $^{\pm0.67}$ | 74.85 $^{\pm0.43}$ | 71.05 $^{\pm0.9}$ | **75.13** $^{\pm0.31}$ |
| | | ven | 64.06 $^{\pm0.31}$ | 63.11 $^{\pm1.27}$ | 64.39 $^{\pm0.36}$ | 63.24 $^{\pm1.26}$ | **65.42** $^{\pm0.76}$ |
| | | xho | 70.77 $^{\pm2.45}$ | 68.54 $^{\pm1.44}$ | 72.37 $^{\pm0.39}$ | 67.00 $^{\pm1.27}$ | **72.92** $^{\pm0.29}$ |
| | | zul | 69.44 $^{\pm0.62}$ | 67.74 $^{\pm1.46}$ | 70.28 $^{\pm0.49}$ | 67.17 $^{\pm0.15}$ | **71.20** $^{\pm0.44}$ |
| | Wikiann | amh | 57.76 $^{\pm0.45}$ | 33.96 $^{\pm1.83}$ | 64.27 $^{\pm1.91}$ | 60.16 $^{\pm2.83}$ | **68.11** $^{\pm1.75}$ |
| | | ibo | 73.6 $^{\pm1.32}$ | 70.83 $^{\pm1.86}$ | 73.93 $^{\pm1.12}$ | **76.14** $^{\pm1.42}$ | 75.73 $^{\pm2.78}$ |
| | | kin | 69.67 $^{\pm2.07}$ | 77.35 $^{\pm4.47}$ | **82.24** $^{\pm2.17}$ | 79.8 $^{\pm1.06}$ | 79.78 $^{\pm1.78}$ |
| | | swh | 88.09 $^{\pm0.32}$ | 88.00 $^{\pm0.28}$ | 88.83 $^{\pm0.47}$ | 86.13 $^{\pm0.2}$ | **89.16** $^{\pm0.35}$ |
| | | yor | 83.8 $^{\pm2.06}$ | 81.96 $^{\pm0.88}$ | **87.96** $^{\pm1.24}$ | 82.77 $^{\pm0.23}$ | 85.00 $^{\pm2.42}$ |
| Phrase Chunking | phrase-chunk | afr | 95.34 $^{\pm0.16}$ | 95.68 $^{\pm0.30}$ | 95.13 $^{\pm0.06}$ | 90.22 $^{\pm0.81}$ | **96.01** $^{\pm0.14}$ |
| | | nso | 96.57 $^{\pm0.61}$ | 96.85 $^{\pm0.55}$ | **98.36** $^{\pm0.2}$ | 96.47 $^{\pm0.14}$ | 98.28 $^{\pm0.1}$ |
| | | sot | 82.93 $^{\pm0.38}$ | 83.08 $^{\pm0.78}$ | 85.28 $^{\pm0.61}$ | 82.18 $^{\pm0.93}$ | **85.69** $^{\pm0.76}$ |
| | | ssw | 82.9 $^{\pm1.03}$ | 81.91 $^{\pm0.47}$ | **84.73** $^{\pm0.18}$ | 83.24 $^{\pm0.11}$ | 83.45 $^{\pm0.12}$ |
| | | tsn | 92.77 $^{\pm0.16}$ | 92.64 $^{\pm0.66}$ | 94.11 $^{\pm0.49}$ | 92.71 $^{\pm0.42}$ | **94.03** $^{\pm0.19}$ |
| | | tso | 86.42 $^{\pm0.46}$ | 86.90 $^{\pm0.31}$ | 87.39 $^{\pm0.18}$ | 86.73 $^{\pm0.95}$ | **89.32** $^{\pm0.43}$ |
| | | ven | 92.31 $^{\pm0.45}$ | 90.47 $^{\pm0.32}$ | 92.42 $^{\pm0.68}$ | 92.02 $^{\pm0.33}$ | **92.54** $^{\pm0.21}$ |
| | | zul | 87.30 $^{\pm0.26}$ | 87.29 $^{\pm1.04}$ | 88.67 $^{\pm0.66}$ | 85.74 $^{\pm0.55}$ | **90.05** $^{\pm0.81}$ |

Table E.4: Performance of mPLMs on each language in each task. ($F_1$) score is the evaluation metric. We use **Red** highlights to indicate languages in zero-shot setting.

| ISO-639-3 | Language | ISO-639-3 | Language | ISO-639-3 | Language | ISO-639-3 | Language |
|---|---|---|---|---|---|---|---|
| aar | Afar / Qafar | bky | Bokyi | dow | Doyayo | gol | Gola |
| aba | Abe / Abbey | bmo | Bambalang | dsh | Daasanach | gqr | Gor |
| abn | Abua | bmv | Bum | dua | Douala | gso | Gbaya, Southwest |
| acd | Gikyode | bom | Berom | dug | Chiduruma | gud | Dida, Yocoboue |
| ach | Acholi | bov | Tuwuli | dwr | Dawro | gur | Farefare |
| ada | Dangme | box | Bwamu / Buamu | dyi | Sénoufo, Djimini | guw | Gun |
| adh | Jopadhola / Adhola | bqc | Boko | dyu | Jula | gux | Gourmanchema |
| adj | Adjukru / Adioukrou | bqj | Bandial | ebr | Ebrie | guz | Ekegusii |
| afr | Afrikaans | bsc | Oniyan | ebu | Kiembu / Embu | gvl | Gulay |
| agq | Aghem | bsp | Baga Sitemu | efi | Efik | gwr | Gwere |
| aha | Ahanta | bss | Akoose | ego | Eggon | gya | Gbaya, Northwest |
| ajg | Aja | bst | Basketo | eka | Ekajuk | hag | Hanga |
| akp | Siwu | bud | Ntcham | eko | Koti | har | Harari |
| alz | Alur | bum | Bulu | eto | Eton | hau | Hausa |
| amh | Amharic | bun | Sherbro | etu | Ejagham | hay | Haya |
| ann | Obolo | bus | Bokobaru | etx | Iten / Eten | hbb | Nya huba |
| anu | Anyuak / Anuak | buy | Bullom So | ewe | Ewe | heh | Hehe |
| anv | Denya | bwr | Bura Pabir | ewo | Ewondo | her | Herero |
| asa | Asu | bwu | Buli | fak | Fang | hgm | Haillom |
| asg | Cishingini | bxk | Bukusu | fat | Fante | hna | Mina |
| atg | Ivbie North-Okpela-Arhe | byf | Bete | ffm | Fulfulde, Maasina | ibb | Ibibio |
| ati | Attie | byv | Medumba | fia | Nobiin | ibo | Igbo |
| avn | Avatime | bza | Bandi | fip | Fipa | idu | Idoma |
| avu | Avokaya | bzw | Basa | flr | Fuliiru | igb | Ebira |
| azo | Awing | cce | Chopi | fon | Fon | ige | Igede |
| bam | Bambara | chw | Chuabo | fub | Fulfulde, Adamawa | igl | Igala |
| bav | Vengo | cjk | Chokwe | fue | Fulfulde, Borgu | ijn | Kalabari |
| bba | Baatonum | cko | Anufo | fuf | Pular | ikk | Ika |
| bbj | Ghomala | cme | Cerma | fuh | Fulfulde, Western Niger | ikw | Ikwere |
| bbk | Babanki | cop | Coptic | ful | Fulah | iqw | Ikwo |
| bci | Baoule | cou | Wamey | fuq | Fulfulde Central Eastern Niger | iri | Rigwe |
| bcn | Bali | crs | Seychelles Creole | fuv | Fulfude Nigeria | ish | Esan |
| bcw | Bana | csk | Jola Kasa | gaa | Ga | iso | Isoko |
| bcy | Bacama | cwe | Kwere | gax | Oromo, Borana-Arsi-Guji | iyx | yaka |
| bdh | Baka | daa | Dangaleat | gaz | Oromo, West Central | izr | Izere |
| bds | Burunge | dag | Dagbani | gbo | Grebo, Northern | izz | Izii |
| bem | Bemba / Chibemba | dav | Dawida / Taita | gbr | Gbagyi | jgo | Ngomba |
| beq | Beembe | dga | Dagaare | gde | Gude | jib | Jibu |
| ber | Berber | dgd | Dagaari Dioula | gid | Gidar | jit | Jita |
| bex | Jur Modo | dgi | Dagara, Northern | giz | South Giziga | jmc | Machame |
| bez | Bena | dhm | Dhimba | gjn | Gonja | kab | Kabyle |
| bfa | Bari | dib | Dinka, South Central | gkn | Gokana | kam | Kikamba |
| bfd | Bafut | did | Didinga | gkp | Kpelle, Guinea | kbn | Kare |
| bfo | Birifor, Malba | dig | Chidigo | gmv | Gamo | kbo | Keliko |
| bib | Bisa | dik | Dinka, Southwestern | gna | Kaansa | kbp | Kabiye |
| bim | Bimoba | dip | Dinka, Northeastern | gnd | Zulgo-gemzek | kby | Kanuri, Manga |
| bin | Edo | diu | Gciriku | gng | Ngangam | kcg | Tyap |
| biv | Birifor, Southern | dks | Dinka, Southeastern | gof | Goofa | kck | Kalanga |
| bjv | Bedjond | dnj | Dan | gog | Gogo | kdc | Kutu |

Table F.1: Languages covered in SERENGETI - Part I.

| ISO-639-3 | Language | ISO-639-3 | Language | ISO-639-3 | Language | ISO-639-3 | Language |
|---|---|---|---|---|---|---|---|
| kde | Makonde | laj | Lango | mfh | Matal | ngb | Ngbandi, Northern |
| kdh | Tem | lam | Lamba | mfi | Wandala | ngc | Ngombe |
| kdi | Kumam | lap | Laka | mfk | Mofu, North | ngl | Lomwe |
| kdj | Ng'akarimojong | lee | Lyélé | mfq | Moba | ngn | Bassa |
| kdl | Tsikimba | lef | Lelemi | mfz | Mabaan | ngo | Ngoni |
| kdn | Kunda | lem | Nomaande | mgc | Morokodo | ngp | Ngulu |
| kea | Kabuverdianu | lgg | Lugbara | mgh | Makhuwa-Meetto | nhr | Naro |
| ken | Kenyang | lgm | Lega-mwenga | mgo | Meta' | nhu | Noone |
| khy | Kele / Lokele | lia | Limba, West-Central | mgq | Malila | nih | Nyiha |
| kia | Kim | lik | Lika | mgr | Mambwe-Lungu | nim | Nilamba / kinilyamba |
| kik | Gikuyu / Kikuyu | lin | Lingala | mgw | Matumbi | nin | Ninzo |
| kin | Kinyarwanda | lip | Sekpele | mif | Mofu-Gudur | niy | Ngiti |
| kiz | Kisi | lmd | Lumun | mkl | Mokole | nka | Nkoya / ShiNkoya |
| kki | Kagulu | lmp | Limbum | mlg | Malagasy | nko | Nkonya |
| kkj | Kako | lnl | Banda, South Central | mlr | Vame | nla | Ngombale |
| kln | Kalenjin | log | Logo | mmy | Migaama | nnb | Nande / Ndandi |
| klu | Klao | lom | Loma | mnf | Mundani | nnh | Ngiemboon |
| kma | Konni | loq | Lobala | mnk | Mandinka | nnq | Ngindo |
| kmb | Kimbundu | lot | Latuka | moa | Mwan | nse | Chinsenga |
| kmy | Koma | loz | Silozi | mos | Moore | nnw | Nuni, Southern |
| knf | Mankanya | lro | Laro | moy | Shekkacho | nso | Sepedi |
| kng | Kongo | lsm | Saamya-Gwe / Saamia | moz | Mukulu | ntr | Delo |
| knk | Kuranko | lth | Thur / Acholi-Labwor | mpe | Majang | nuj | Nyole |
| kno | Kono | lto | Tsotso | mpg | Marba | nus | Nuer |
| koo | Konzo | lua | Tshiluba | mqb | Mbuko | nwb | Nyabwa |
| koq | Kota | luc | Aringa | msc | Maninka, Sankaran | nxd | Ngando |
| kqn | Kikaonde | lue | Luvale | mur | Murle | nya | Chichewa |
| kqp | Kimré | lug | Luganda | muy | Muyang | nyb | Nyangbo |
| kqs | Kisi | lun | Lunda | mwe | Mwera | nyd | Olunyole / Nyore |
| kqy | Koorete | luo | Dholuo / Luo | mwm | Sar | nyf | Giryama |
| kri | Krio | lwg | Wanga | mwn | Cinamwanga | nyk | Nyaneka |
| krs | Gbaya | lwo | Luwo | mws | Mwimbi-Muthambi | nym | Nyamwezi |
| krw | Krahn, Western | maf | Mafa | myb | Mbay | nyn | Nyankore / Nyankole |
| krx | Karon | mas | Maasai | myk | Sénoufo, Mamara | nyo | Nyoro |
| ksb | Shambala / Kishambala | maw | Mampruli | myx | Masaaba | nyu | Nyungwe |
| ksf | Bafia | mbu | Mbula-Bwazza | mzm | Mumuye | nyy | Nyakyusa-Ngonde / Kyangonde |
| ksp | Kabba | mck | Mbunda | mzw | Deg | nza | Mbembe, Tigon |
| ktj | Krumen, Plapo | mcn | Masana / Massana | naq | Khoekhoe | nzi | Nzema |
| ktu | Kikongo | mcp | Makaa | naw | Nawuri | odu | Odual |
| kua | Oshiwambo | mcu | Mambila, Cameroon | nba | Nyemba | ogo | Khana |
| kub | Kutep | mda | Mada | nbl | IsiNdebele | oke | Okpe |
| kuj | Kuria | mdm | Mayogo | ncu | Chunburung | okr | Kirike |
| kus | Kusaal | mdy | Maale | ndc | Ndau | oku | Oku |
| kvj | Psikye | men | Mende | nde | IsiNdebele | orm | Oromo |
| kwn | Kwangali | meq | Merey | ndh | Ndali | ozm | Koonzime |
| kyf | Kouya | mer | Kimiiru | ndj | Ndamba | pcm | Nigerian Pidgin |
| kyq | Kenga | mev | Maan / Mann | ndo | Ndonga | pem | Kipende |
| kzr | Karang | mfe | Morisyen / Mauritian Creole | ndv | Ndut | pkb | Kipfokomo / Pokomo |
| lai | Lambya | mfg | Mogofin | ndz | Ndogo | | |

Table F.2: Languages covered in SERENGETI - Part II

| ISO-639-3 | Language | ISO-639-3 | Language | ISO-639-3 | Language |
|---|---|---|---|---|---|
| pov | Guinea-Bissau Creole | tcd | Tafi | won | Wongo |
| poy | Pogolo / Shipogoro-Pogolo | ted | Krumen, Tepo | xan | Xamtanga |
| rag | Lulogooli | tem | Timne | xed | Hdi |
| rel | Rendille | teo | Teso | xho | Isixhosa |
| rif | Tarifit | tex | Tennet | xnz | Mattokki |
| rim | Nyaturu | tgw | Senoufo, Tagwana | xog | Soga |
| rnd | Uruund | thk | Tharaka | xon | Konkomba |
| rng | Ronga / ShiRonga | thv | Tamahaq, Tahaggart | xpe | Kpelle |
| rub | Gungu | tir | Tigrinya | xrb | Karaboro, Eastern |
| run | Rundi / Kirundi | tiv | Tiv | xsm | Kasem |
| rwk | Rwa | tke | Takwane | xtc | Katcha-Kadugli-Miri |
| sag | Sango | tlj | Talinga-Bwisi | xuo | Kuo |
| saq | Samburu | tll | Otetela | yal | Yalunka |
| sba | Ngambay | tog | Tonga | yam | Yamba |
| sbd | Samo, Southern | toh | Gitonga | yao | Yao / Chiyao |
| sbp | Sangu | toi | Chitonga | yat | Yambeta |
| sbs | Kuhane | tpm | Tampulma | yba | Yala |
| sby | Soli | tsc | Tshwa | ybb | Yemba |
| sef | Sénoufo, Cebaara | tsn | Setswana | yom | Ibinda |
| ses | Songhay, Koyraboro Senni | tso | Tsonga | yor | Yoruba |
| sev | Sénoufo, Nyarafolo | tsw | Tsishingini | yre | Yaoure |
| sfw | Sehwi | ttj | Toro / Rutoro | zaj | Zaramo |
| sgw | Sebat Bet Gurage | ttq | Tawallammat | zdj | Comorian, Ngazidja |
| shi | Tachelhit | ttr | Nyimatli | zga | Kinga |
| shj | Shatt | tui | Toupouri | ziw | Zigula |
| shk | Shilluk | tul | Kutule | zne | Zande / paZande |
| sid | Sidama | tum | Chitumbuka | zul | Isizulu |
| sig | Paasaal | tuv | Turkana | | |
| sil | Sisaala, Tumulung | tvu | Tunen | | |
| sna | Shona | twi | Twi | | |
| snf | Noon | umb | Umbundu | | |
| sng | Sanga / Kiluba | urh | Urhobo | | |
| snw | Selee | uth | ut-Hun | | |
| som | Somali | vag | Vagla | | |
| sop | Kisonge | vai | Vai | | |
| sor | Somrai | ven | Tshivenda | | |
| sot | Sesotho | vid | Chividunda | | |
| soy | Miyobe | vif | Vili | | |
| spp | Senoufo, Supyire | vmk | Makhuwa-Shirima | | |
| ssw | Siswati | vmw | Macua | | |
| suk | Sukuma | vun | Kivunjo | | |
| sus | Sosoxui | vut | Vute | | |
| swa | Swahili | wal | Wolaytta | | |
| swc | Swahili Congo | wbi | Vwanji | | |
| swh | Swahili | wec | Guere | | |
| swk | Sena, Malawi | wes | Pidgin, Cameroon | | |
| sxb | Suba | wib | Toussian, Southern | | |
| taq | Tamasheq | wmw | Mwani | | |
| tcc | Datooga | wol | Wolof | | |

Table F.3: Languages covered in SERENGETI - Part III.

| ISO-639-3 | #Tokens |
|---|---|
| swh | 2,912,488,735 |
| afr | 1,264,478,436 |
| som | 587,549,878 |
| swa | 499,792,448 |
| hau | 286,806,539 |
| amh | 241,700,000 |
| mlg | 137,852,716 |
| zne | 89,981,183 |
| sna | 75,413,519 |
| ... | ... |
| bam | 3,262 |
| har | 3,066 |
| dyo | 1,797 |
| fvr | 1,680 |
| tbz | 1,578 |
| ddn | 1,372 |
| fuc | 1,336 |
| knc | 1,097 |
| eot | 1,041 |
| cgg | 845 |

Table F.4: The sizes of the top 10 and bottom 10 languages in SERENGETI pretraining.

## ACL 2023 Responsible NLP Checklist

### A For every submission:

☑ A1. Did you describe the limitations of your work?
*8*

☑ A2. Did you discuss any potential risks of your work?
*8*

☑ A3. Do the abstract and introduction summarize the paper's main claims?
*1 and 7*

☒ A4. Have you used AI writing assistants when working on this paper?
*Left blank.*

### B ☑ Did you use or create scientific artifacts?

*4 and 5*

☑ B1. Did you cite the creators of artifacts you used?
*2, 4 and 5*

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
*2, 4 and 5*

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
*9*

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
*We use only publicly available data to develop our models. Our data comes from 517 languages and language varieties and hence it is challenging to carry out manual investigation on it. However, since the data belong to the public domain, we do not have serious concerns about privacy or anti-social language beyond what already exists online and is accessible to anyone.*

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
*2, 3, 4, 5, 6*

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
*2, 3, 4, 5, 6*

*The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.*

**C** ☑ **Did you run computational experiments?**

*5 and 6*

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
*5*

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
*4, 5*

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
*2, 3, 4, 5, 6, Appendix*

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
*2, 4, 5, 6, Appendix*

**D** ☑ **Did you use human annotators (e.g., crowdworkers) or research with human participants?**

*7*

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
*Not applicable. Left blank.*

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
*Not applicable. Left blank.*

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
*Not applicable. Left blank.*

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
*Not applicable. Left blank.*

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
*Not applicable. Left blank.*