AfroLM: A Self-Active Learning-based Multilingual Pretrained Language Model for 23 African Languages

Bonaventure F. P. Dossou[1,2,*], **Atnafu Lambebo Tonja**[3,*], **Oreen Yousuf**[4,*], **Salomey Osei**[5,*], **Abigail Oppong**[6,*], **Iyanuoluwa Shode**[7,*], **Oluwabusayo Olufunke Awoyomi**[8,*], **Chris Chinenye Emezue**[1,9,*]

[*]Masakhane NLP, [1]Mila Quebec AI Institute, Canada, [2] McGill University, Canada, [3]Instituto Politécnico Nacional, Mexico, [4]Uppsala University, Sweden, [5]Universidad de Deusto, Spain [6]Ashesi University, Ghana, [7]Montclair State University, USA, [8]The College of Saint Rose, USA, [9]Technical University of Munich, Germany

Abstract

In recent years, multilingual pre-trained language models have gained prominence due to their remarkable performance on numerous downstream Natural Language Processing tasks (NLP). However, pre-training these large multilingual language models requires a lot of training data, which is not available for African Languages. Active learning is a semi-supervised learning algorithm, in which a model consistently and dynamically learns to identify the most beneficial samples to train itself on, in order to achieve better optimization and performance on downstream tasks. Furthermore, active learning effectively and practically addresses real-world data scarcity. Despite all its benefits, active learning, in the context of NLP and especially multilingual language models pretraining, has received little consideration. In this paper, we present **AfroLM**, a multilingual language model pretrained from scratch on 23 African languages (the largest effort to date) using our novel self-active learning framework. Pretrained on a dataset significantly (14x) smaller than existing baselines, **AfroLM** outperforms many multilingual pretrained language models (AfriBERTa, XLMR-base, mBERT) on various NLP downstream tasks (NER, text classification, and sentiment analysis). Additional out-of-domain sentiment analysis experiments show that **AfroLM** is able to generalize well across various domains. We release the code source, and our datasets used in our framework at `https://github.com/bonaventuredossou/MLM_AL`.

1 Introduction

With the appearance of Transformer models (Vaswani et al., 2017), the field of Natural Language Processing (NLP) has seen the emergence of powerful multilingual pre-trained language models (MPLMs), such as mBERT (Devlin et al., 2018), XLM-RoBERTa (XML-R) (Conneau et al., 2019), and mT5 (Xue et al., 2021). These prominent models have helped achieve state-of-the-art (SOTA) performance in many downstream NLP tasks such as named entity recognition (NER) (Alabi et al., 2022a; Adelani et al., 2021a; Devlin et al., 2018; Conneau et al., 2019), text classification (Kelechi et al., 2021), and sentiment analysis (Alabi et al., 2022a; Adelani et al., 2021a; Devlin et al., 2018; Conneau et al., 2019). However they usually require a large amount of unlabeled text corpora for good performance: mBERT was trained on Wikipedia (2,500M words) and BookCorpus (Zhu et al., 2015) (800M words) across 104 languages - 5 of which are African; mT5 supports 101 languages (13 African) and XLM-R supports 100 languages (8 African), and were trained on mC4 (Xue et al., 2021) and CommonCrawl data (Wenzek et al., 2019), respectively. This requirement for large-scale datasets contrasts sharply with the scarcity of available text corpora for African languages, which has pushed them into low-resource settings and largely excluded them from the pretraining phase of these large pre-trained models (Joshi et al., 2020; Adelani et al., 2022a). This exclusion, leads very often, to a poor performance on languages unseen during pre-training (Alabi et al., 2022a) which eventually leads to inability to carry out the required NLP task.

Active learning is a semi-supervised machine learning algorithm that makes use of only a few initial training data points to achieve better performance of a given model **M**. The optimization is done by iteratively training **M**, and using another model **N**, usually referred to as the *oracle*, to choose new training samples that will help **M** find better configurations while improving its performance (e.g., prediction accuracy). This makes active learning a prevalent paradigm to cope with data scarcity. The efficiency of active learning (i.e. its ability to produce better performance despite being trained on a smaller training data) has been proven in tasks such as biological sequence design (Jain et al., 2022), chemical sampling (Smith

Proceedings of The Third Workshop on Simple and Efficient Natural Language Processing (SustaiNLP), pages 52 - 64
December 7, 2022 ©2022 Association for Computational Linguistics

Languages	Family	Writing System	African Region	No of Speakers	Initial # of Sentences	Source	Size (MB)
Amharic (amh)	Afro-Asiatic/Semitic	Ge'ez script	East	57M	655,079	✣,✝,★	279
Afan Oromo (orm)	Afro-Asiatic/Cushitic	Latin script	East	37.4M	50,105	✝	9.87
Bambara (bam)	NC/Manding	Latin, Arabic(Ajami), N'ko	West	14M	6,618	✣	1.00
Ghomálá' (bbj)	NC/Grassfields	Latin script	Central	1M	4,841	✣	0.50
Éwé (ewe)	NC/Kwa	Latin (Ewe alphabet)	West	7M	5,615	✣	0.50
Fon (fon)	NC/Volta-Niger	Latin script	West	1.7M	5,448	✣	1.00
Hausa (hau)	Afro-Asiatic/Chadic	Latin (Boko alphabet)	West	63M	1,626,330	✣,✝,★	208
Igbo (ibo)	NC/Volta-Niger	Latin (Önwu alphabet)	West	27M	437,737	✣,✝,★	63
Kinyarwanda (kin)	NC/Rwanda-Rundi	Latin script	Central	9.8M	84,994	➼,✝,✣	37.70
Lingala (lin)	NC/Bang	Latin script	Central & East	45M	398,440	✣	45.90
Luganda (lug)	NC/Bantu	Latin script (Ganda alphabet)	East	7M	74,754	✝,✣	8.34
Luo (luo)	Nilo-Saharan	Latin script	East	4M	8,684	✝	1.29
Mooré (mos)	NC/Gur	Latin script	West	8M	27,908	✣,✝	5.05
Chewa (nya)	NC/Nyasa	Latin script	South & East	12M	8,000	✣	1.66
Naija (pcm)	English-Creole	Latin script	West	75M	345,694	✣,✝,★	101
Shona (sna)	NC/Bantu	Latin script (Shona alphabet)	Southeast	12M	187,810	✣,✝	32.80
Swahili (swa)	NC / Bantu	Latin script (Roman Swahili alphabet)	East & Central	98M	1,935,485	✣,✝,★	276
Setswana (tsn)	NC / Bantu	Latin (Tswana alphabet)	South	14M	13,958	✣,✝	2.21
Akan/Twi (twi)	NC / Kwa	Latin script	West	9M	14,701	✣	1.61
Wolof (wol)	NC / Senegambia	Latin (Wolof alphabet)	West	5M	13,868	✝	2.20
Xhosa (xho)	NC/Zunda	Latin (Xhosa alphabet)	South	20M	93,288	✣,✝	17.40
Yorůbá (yor)	NC / Volta-Niger	Latin (Yorùbá alphabet)	West	42M	290,999	✣,✝,★	45.9
isiZulu (zul)	NC / Bantu	Latin (Zulu alphabet)	South	27M	194,562	✣,✝	33.70

Table 1: **Languages Corpora Details**. **Legends**: (Adelani et al., 2022a) → ✣, (Alabi et al., 2022a) → ✝, (Kelechi et al., 2021) → ★, (Niyongabo et al., 2020) → ➼.

et al., 2018), and Deep Bayesian (DB) approaches on image data (Gal et al., 2017). Also, most of the work on deep active learning focuses on image classification with Convolutional Neural Networks (CNNs). It should be noted that active learning has been greatly explored and used to perform classification tasks, but not in language generation and understanding, and this is what we hope to address.

A study of active learning in the context of NLP has been carried out by (Siddhant and Lipton, 2018). In their study, it is shown that active learning with DB networks coupled with uncertainty measures and acquisition function outperforms several i.i.d baselines. They showed that with only 20% of samples labeled, their approach reached an accuracy of 98-99% on the Named Entity Recognition (NER) task, while i.i.d tasks required 50% of labelled data to achieve comparable performance. In their study on clinical texts, (Chen et al., 2015) also proved that active learning algorithms outperformed other learning methods. (Ein-Dor et al., 2020; Tonneau et al., 2022) on their works with BERT model(s) (for n different languages, there were n different BERT-based models) went further by showing that active learning works with a balanced and unbalanced dataset. They also showed that the different active learning methods performed relatively the same.

In our work, we fixed **M=N** (hence the title **self-active learning**). In our framework, we give **M** the ability to query itself, and use the knowledge acquired during each active learning round to construct new data points (from existing ones) that will be used for the next active learning round.

We considered a diverse set of 23 African languages spread across the African continent. The selected languages are spoken in the south, central, east, and western regions of Africa. The languages cover four language families: Afro-Asiatic (e.g., Amharic, Hausa, Afan Omoro), Niger-Congo (NC) (e.g., Yorùbá, Bambara, Fon), English-Creole (Naija) and Nilo-Saharan (Luo) (see Appendix A for details). For each language, a dataset was collected from the news domain, which encompassed many topics such as health, politics, society, sport, environment, etc.

Our primary contribution to this work is our proposal of a **self-active learning framework** in which we pre-train the **biggest Multilingual African Language Model** (for the number of languages covered) to date, and we show that our setup is **very data-efficient and provides improvements** on downstream NLP tasks such as NER, text classification, and sentiment analysis (even on out-of-domain experiments).

2 Related Works on MPLMs for African Languages

Language adaptive fine-tuning (LAFT) is one of the best approaches to adapt MPLMs to a new language. This entails fine-tuning an MPLM on monolingual texts of the said language with the same pre-training objective. However, this cannot be efficiently applied to African languages facing data-scarcity. (Alabi et al., 2022b) proposed a new adaptation method called *Multilingual adaptive fine-tuning (MAFT)*, as an approach to adapt MPLMs to many African languages with a single

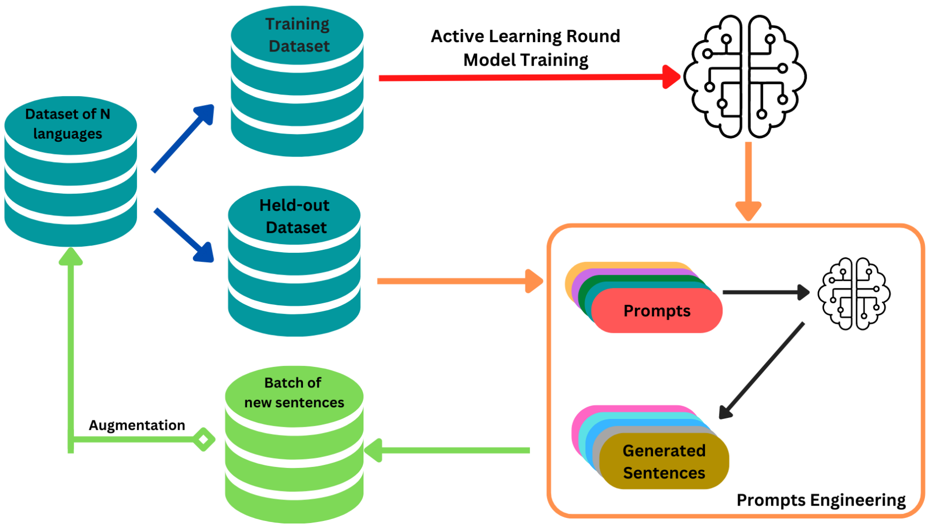

Figure 1: Self-Active Learning Framework). The process is designed in 4 stages (fully explained and detailed in Algorithm 1): (1) 🟦 Dataset split for current Active Learning round, (2) 🟥 Active Learning round training, (3) 🟧 Generation of new sentence samples for the current round, and (4) 🟩 Augmentation of the datasets of all languages.

model. Their results show that MAFT is competitive to LAFT while providing a single model rather than many models that are specific for individual languages. Nevertheless, Alabi et al. (2022b)'s approach still works under the assumption that one does not need to train a model from scratch for languages in the low-resource settings, as they could benefit from high-resource languages. We find that this is not *always* the case.

(Kelechi et al., 2021) introduced AfriBERTa, a multilingual language model trained on less than 1GB of data from 11 African languages. Training AfriBERTa from scratch showcased how African languages can benefit from being included in the pre-training stage of MPLMs. AfriBERTa produced competitive results compared to existing MPLMs (e.g., mBERT, XLM-R), and outperformed them on text classification and NER tasks. Rather than relying on high-resource languages for transfer-learning, AfriBERTa leverages the linguistic similarity between languages with low-resource settings to produce promising results. (Kelechi et al., 2021) empirically demonstrates that this is more beneficial to these languages and is crucial in assessing the viability of language models pre-trained on small datasets.

(Antoine and Niyongabo, 2022) went beyond the linguistic taxonomy in creating KinyaBERT, a morphology-aware language model for Kinyarwanda. Trained on a 2.4GB corpus containing news articles from 370 websites registered between 2011 and 2021, KinyaBERT boasts a Transformer-like architecture that helps the representation of morphological compositionality. Their experiments outperformed solid baseline results for tasks such as NER and machine-translated GLUE on the Kinyarwanda language. These results demonstrated the effectiveness of not relying on transfer learning from high resource languages and rather explicitly incorporating morphological information of the African languages in their pre-training stage.

In the next section, we will describe our self-active learning framework, and the core details of our approach.

3 Self-Active Learning Framework

In this section, we describe our self-active learning framework (Figure 1). In Algorithm 1, we present a single active learning loop. In our current work, our model is trained only with a Masked Language Modeling (MLM) objective (Conneau et al., 2019; Conneau and Lample, 2019; Devlin et al., 2018). We plan to further incorporate Translation Language Modeling (TLM) objective to improve translations of low-resource languages with relatively few thousands of data points [1]. This will be useful for both supervised and unsupervised translation (Adelani et al., 2022a; Conneau et al., 2019).

We used a shared Sentence Piece vocabulary with $250,000$ BPE codes. The subword shared

[1] https://github.com/facebookresearch/XLM

vocabulary intends to improve alignment in the embedding space across languages (see languages description in Appendix A and corpora details in Table 1) that are linguistically similar in features such as script/alphabet, morphology, etc. (Conneau et al., 2019), reflecting our focus languages. Additionally, (Conneau et al., 2019) showed that scaling the size of the shared vocabulary (e.g. from 36,000 to 256,000) improved the performance of multilingual models on downstream tasks. Our vocabulary is defined jointly across all 23 languages and fixed during training, as opposed to random training and held-out dataset selection at each active learning round.

The motivations behind the randomness in the selection of the training and held-out datasets are: (1) to make efficient use of the limited dataset we have, and (2) to expose the model step by step, instead of simultaneously, to a variety of samples across different news sub-domains. We believe this would help in domain-shift adaptation and the robustness of the model.

As extensively detailed in Algorithm 1, at each round we randomly select m sentences per language, from the held-out dataset of the language. For a language, to generate a new sentence s', given an original sentence s, we proceed as follows (more details can be found in Algorithm 1):

1. select an initial ordered (left to right) set of words from s as prompt,
2. add a mask token at the end of the ordered set or sequence of words,
3. query the model to predict the masked token,
4. choose the best word, add it to the prompt,
5. repeat 2-4 until we reach the length of s.

The process described above will produce m new data points that will be added to the language dataset. The new dataset obtained is used to re-train the model from scratch at the next active learning round.

4 Experiments, Results and Discussion

Experiments: We use the XLM-RoBERTa (XLM-R) architecture in our experiments based on previous works utilizing the model to achieve state-of-the-art performance in various downstream tasks. Following the work and results of (Kelechi et al., 2021), we trained XLM-R-based models

Algorithm 1 Self-Active Learning Training Round

```
Require:
  • Masked Language Modeling (MLM) objective $\pi_\theta$ with masking probability $p = 0.15$
  • Vocabulary $\mathcal{V}$, Model $\mathcal{M}$, Tokenizer $\mathcal{T}$
  • Set of languages $\mathcal{L} = \bigcup_{i \in [1,23]} \{l\}$
  • Overall Dataset $\mathcal{D} = \bigcup_{l \in \mathcal{L}} \mathcal{D}_l$ with $\mathcal{D}_l$ the dataset of language $l$
  • Training Dataset $\mathcal{D}_t$ with $k\%$ randomly selected sentences from $\mathcal{D}_l, l \in \mathcal{L}$
  • Held-out Dataset $\mathcal{H}$ with $1 - k\%$ samples for each language: $\mathcal{H} = \bigcup_{l \in L} \mathcal{H}_l$
  • proportion $t$ of words to successively mask in a sentence (from left to right)
Ensure:
  • Initialize $\mathcal{M}$, and $\mathcal{T}$ with $\mathcal{V}$
  • $k \leftarrow 80$
  • $t \leftarrow 15$
  • Train $\mathcal{M}$ with policy $\pi_\theta$
  Generate set $\mathcal{G}_l$ of new samples for each language:
  for $l \in L$ do
      $\mathcal{G}_l \leftarrow \{\}$
      • Build $\mathcal{S}_l$ with $m = |\mathcal{H}_l|$ sentences randomly chosen from $\mathcal{H}_l$   ▷ we choose $m$ this way to cope with small size datasets
      for $s \in \mathcal{S}_l$ do
          $n \leftarrow len(s)$, $s = \bigcup_{i \in [1,n]} \{w_i\}$
          $t_s \leftarrow \lceil \frac{n*t}{100} \rceil + 1$
          $prompt \leftarrow \bigcup_{i \in [1,n-t_s]} \{w_i\}$
          while $t_s \neq 0$ do
              $prompt \leftarrow prompt \cup \{$<mask>$\}$
              $w_p \leftarrow \mathcal{M}(prompt)$:   ▷ predicted masked word
              $prompt \leftarrow prompt \cup \{w_p\}$
              $t_s \leftarrow t_s - 1$
          end while
          $\mathcal{G}_l \leftarrow \mathcal{G}_l \cup \{prompt\}$
      end for
      $\mathcal{D}_l \leftarrow \mathcal{D}_l \cup \mathcal{G}_l$   ▷ new samples added to the language dataset
  end for
```

Model	Hyper-parameters	Values
AfroLM-Large	sequence maximum length	256
	hidden size	768
	attention heads	6
	hidden layers	10
	learning rate	1e-4
	batch size	32
	# of Parameters	264M
	total initial training examples	5,137,026
	vocabulary size	250,000
	gradient accumulation steps	8
	warming steps	40,000
	training steps	500,000

Table 2: Hyper-parameters summary

from scratch. In our current work we trained the model with 3 self-active learning rounds (we stopped at 3 due to computational resources). We used 80% and 20% of languages data for the training and held-out datasets respectively. We designed 2 versions of AfroLM: *AfroLM-Large (without self-active learning)* and *AfroLM-Large (with self-active learning)* with the hyper-parameters specified in Table 2. All training experiments were done using the HuggingFace Transformers library (Wolf et al., 2019).

Afro (without self-active learning) is one of our baselines. We trained an XLM-R model on the entire dataset, and the held-out dataset was just used for evaluation. For **AfroLM-Large** models, we used Google Cloud with a single 48GB NVIDIA A100 GPU. An active learning round took ≈ 260 hours of training. We evaluated **AfroLM-Large** models on three downstream tasks:

- **NER**: we evaluated the performance of our model pre-trained using our self-active learning framework on the MasakhaNER dataset (Adelani et al., 2021a). The dataset contains ten African languages: Amharic, Hausa, Igbo, Kinyarwanda, Luganda, Luo, Nigerian Pidgin, Swahili, Wolof, and Yorùbá. (Adelani et al., 2021a; Alabi et al., 2022a) also provided strong baselines with pre-trained language models like mBERT and XLM-R on MasakhaNER.

- **Text Classification**: we tested our models on Hausa and Yorùbá news text classification dataset from (Hedderich et al., 2020), where the authors have also built strong baselines on mBERT and XLM-R models.

- **Sentiment Analysis**: we tested the the out-of-domain performance of our model in two domains different from news:

  1. `Movies`: we directly fine-tuned and evaluated **AfroLM-Large** on the YOSM dataset (Shode et al., 2022), which contains reviews of Yorùbá movies.
  2. `Twitter` → `Movies`: in this setup, we finetuned on the training and validation set of NaijaSenti (Muhammad et al., 2022), and evaluated on YOSM. NaijaSenti contains human annotated tweets in Hausa, Yoruba, Igbo and Nigerian Pidgin. However, we were not able to evaluate **AfroLM-Large** on it because the authors have not yet released the test set.

Results & Discussion: Tables 1 and 3 show that our framework includes a large variety of African Languages. Table 4, and Table 5 (with 11 additional languages from MasakhaNER 2.0 dataset (Adelani et al., 2022b)) show the results of our method in comparison with other baselines on NER task. We can notice that **AfroLM-Large (w/ AL)** outperforms AfriBERTa-Large, mBERT and XLMR-base (≈ 2.5 TB of data); while being pre-trained on significantly smaller dataset (≈ 0.73 GB (80% of 0.91 GB initial dataset)). AfriBERTa-Large has been pretrained from scratch on 11 African languages, while mBERT and XLMR-base (with existing pretrained weights) were finetuned on the MasakhaNER dataset.

Table 6, and Table 7 show that, on the text classification, and sentiment analysis tasks, our method outperforms many existing baselines. Additionally, out-of-domain experiments and analyses show that our method is robust and provides good results in out-of-domain settings.

While AfroXLMR-base in average, slightly outperforms our approach, it is important to notice that it has been pretrained on a dataset 14x bigger than our set. Furthermore, **AfroLM-Large** has been trained on ≈ 0.73 GB of data (80% of 0.91 GB initial dataset), which is less than the size of the corpus used to train AfriBERTa (0.939 GB). This allows us to confidently affirm that our approach is data-efficient, while being very competitive.

It is important to note that the margin of performance from **AfroLM-Large (w/ AL)** does not come from the fact that it has been trained on more languages. Our results show that **AfroLM-Large (w/ AL)** outperforms models trained on significantly larger datasets and number of languages.

Language	**In AfriBERTa?**	**In AfroLM?**	**In AfroXLMR**	**In mBERT?**	**In XLMR?**
`amh`	✓	✓	✓	✗	✓
`hau`	✓	✓	✓	✗	✓
`ibo`	✓	✓	✓	✗	✗
`kin`	✓	✓	✓	✗	✗
`lug`	✗	✓	✗	✗	✗
`luo`	✗	✓	✗	✗	✗
`pcm`	✓	✓	✓	✗	✗
`swa`	✓	✓	✓	✓	✓
`wol`	✓	✓	✓	✗	✗
`yor`	✓	✓	✓	✓	✗

Table 3: Information about languages included in each language model. We can notice that AfroLM includes the most of them.

Language	**AfriBERTa-Large**	**AfroLM-Large (w/o AL)**	**AfroLM-Large (w/ AL)**	**AfroXLMR-base**	**mBERT**	**XLMR-base**
`amh`	73.82	43.78	**73.84**	*76.10*	00.00	70.96
`hau`	90.17	84.14	**91.09**	*91.10*	87.34	87.44
`ibo`	87.38	80.24	**87.65**	*87.40*	85.11	84.51
`kin`	73.78	67.56	**72.84**	*78.00*	70.98	73.93
`lug`	78.85	72.94	**80.38**	*82.90*	80.56	80.71
`luo`	70.23	57.03	**75.60**	*75.10*	72.65	75.14
`pcm`	85.70	73.23	**87.05**	*89.60*	87.78	87.39
`swa`	87.96	74.89	**87.67**	*88.60*	86.37	87.55
`wol`	61.81	53.58	**65.80**	*67.40*	66.10	64.38
`yor`	81.32	73.23	**79.37**	*82.10*	78.64	77.58
avg	79.10	68.06	**80.13**	*81.90*	71.55	79.16
avg (excl. `amh`)	79.69	70.76	**80.83**	*82.54*	79.50	80.07

Table 4: **NER Performances:** F1-scores on languages test sets after 50 epochs averaged over 5 seeds. These results cover all 4 tags in the MasakhaNER dataset: **PER**, **ORG**, **LOC**, **DATE**. XLM-R and mBERT results obtained from (Adelani et al., 2021b). **AfroLM-Large (w/ AL)** outperforms AfriBERTa, and the initial MasakhaNER baselines. **The bold numbers represent the performance of the model with the lowest pretrained data**. AfroXMLR-base = XLMR-Large + MAFT (Alabi et al., 2022a) with 272M parameters. MAFT gives similar performance to individual LAFT models (Alabi et al., 2022a) (LAFT results in single model per language).

Moreover, the comparison of **AfroLM-Large (w/ AL)** to **AfroLM-Large (w/o AL)** shows a significant improvement in performance, which implies that our self-active learning framework is efficient, and leads to a better performance. This is expected, because the idea of our self-active learning (and of active learning in general) is to have **AfroLM**, to consistently and dynamically, during the training phase, identify the most beneficial sample(s) to learn from in order to boost the performance.

In our current algorithm, a sentence sample is generated by ***iterative next-token prediction***: the generated sentence is the result of the concatenation of each best token. Diversity in sample generation and selection is paramount, and we believe, could improve the performance of our framework. In the limitation section (section 6), we proposed a way of selecting *diverse* sentences (after sentence generation). We also proposed a new weighted loss, that we believe will be more balanced across the entire dataset.

5 Future works and Conclusion

In conclusion, we propose **AfroLM**, a self-active learning-based multilingual language model supporting 23 African Languages; the largest to date. Our language datasets are collected from the news domain and span across different parts of the African continent. Our experimental results on NLP downstream tasks (NER, text classification, and out-of-domain sentiment analysis), prove the data-efficiency of **AfroLM** (as it has been trained on a dataset 14x smaller than its competitors), and its competitiveness as it outperforms many MPLMs (AfriBERTa, mBERT, XLMR-base) while being very competitive to *AfroXLMR-base*. We

Model	bam	bbj	ewe	fon	mos	nya	sna	tsn	twi	xho	zul	AVG
MPLMs pre-trained on from scratch on African Languages												
AfriBERTa-Large	78.60	71.00	86.90	79.90	71.40	88.60	92.40	83.20	75.70	85.00	81.70	81.31
AfroLM-Large (w/ AL)	**80.40**	**72.91**	**88.14**	**80.48**	**72.14**	**90.25**	**94.46**	**85.38**	**77.89**	**87.50**	**86.31**	**83.26**
MPLMs adapted to African Languages												
AfroXLMR-base	*79.60*	*73.30*	*89.20*	*82.30*	*74.40*	*91.90*	*95.70*	*87.70*	*78.90*	*88.60*	*88.40*	*84.55*
mBERT	78.90	60.60	86.90	79.90	71.40	88.60	92.40	86.40	75.70	85.00	81.70	80.68
XLMR-base	78.70	72.30	88.50	81.90	72.70	89.90	93.60	86.10	78.70	87.00	84.60	83.09

Table 5: **NER Baselines on MasakhaNER2.0 (Adelani et al., 2022b)**. We compare MPLMs trained from scratch on African languages, and MPLMs adapted to African Languages. The average of scores are over 5 runs. The bold numbers represent the performance of the model with the lowest pretrained data.

Language	In AfriBERTa?	In AfroLM?	AfriBERTa-Large	AfroLM-Large (w/o AL)	AfroLM-Large (w/ AL)
hau	✓	✓	90.86	85.57	**91.00**
yor	✓	✓	*83.22*	75.30	**82.90**

Table 6: **Text Classification Performances:** F1-scores on the languages test sets. **The bold numbers represent the performance of the model with the lowest pretrained data**.

Models	Yoruba F1-score
AfroLM-Large (w/o AL)	
Movies	83.12
Twitter → Movies	41.28
AfroLM-Large (w/ AL)	
Movies	**85.40**
Twitter → Movies	**68.70**
AfriBERTa-Large	
Movies	82.70
Twitter → Movies	65.90

Table 7: **Out-Of-Domain Sentiment Analysis Performance:** F1-scores on YOSM test set after 20 epochs averaged over 5 seeds. **The bold numbers represent the performance of the model with the lowest pretrained data**.

also show that **AfroLM** is also able to generalize across various domains. For future work, we intend to: (1) explore and understand the relationship between the number of active learning steps and the MPLMs performance on downstream tasks, and (2) integrate a new weighted loss, and more diversity in new data points generation and selection as we explained in the limitation section (see section 6). Our datasets, and source code are publicly available at https://github.com/bonaventuredossou/MLM_AL.

6 Limitations and Approach of Solution

Currently, the loss of the model across the training dataset (across all 23 languages), appears to be the average of the individual (cross-entropy) losses. Due to the disparate sizes of our corpora per language, the training will be biased toward the languages whose sizes predominate the training set. Therefore, we suggest a strategy to re-weight the cross entropy loss per language by the ratio of the size of the dataset for that language to the size of the entire training set:

$$\mathcal{L} = \frac{1}{N}\sum_{l} |\frac{\mathcal{D}_l}{\mathcal{D}}|\mathcal{L}_l$$

where $|\frac{\mathcal{D}_l}{\mathcal{D}}|$ is the weight of the training dataset of the language l, $\mathcal{L}_l$ is the loss of the model on a given language l, and N is the total number of languages (23 in our case). We believe this adjusts well overall loss, by using the right weighted loss of each language, that can be seen as their respective contribution to the general loss.

Another limitation of our current framework is that the samples that are generated from prompts might not be diverse. Given a batch $\mathcal{B}$ of generated samples, and a set $\mathcal{S}$ of initial samples, we want the samples selected to be substantially different from the majority of samples present in $\mathcal{S}$. We think that performing the following two steps will help to ensure this:

1. increase the number of words, in a sentence, to be masked: this implies that the length of the prompt is shortened, and that we provide less (or short) context in the input to our model. Long-range semantics is still a challenge in natural language generation and understanding, and large language models (GPT-2, DialoGPT) have insufficiently learned the effect of distant words on next-token prediction (Malkin et al., 2022). Therefore, we believe that providing a short context will increase the choices of the model and lead to the generation of more various tokens. This has been shown by (Malkin et al., 2022) where they also introduced the *coherence boosting* approach to increase the focus of a language model on a long context.

2. use the Word Error Rate (WER) as a simple diversity measurement. The WER is an adaptation of the Levenshtein distance (also called edit distance), working at the word level instead of the phoneme level. Ideally, we want high WER. Let $W = \bigcup_{i \in [1,t_s]} \{w_i\}$, the set of words from a sentence s that we cut off for the next-token prediction loop described in section 3 and in Algorithm 1. Let $W' = \bigcup_{i \in [1,t_s]} \{w'_i\}$, the set of words predicted by the model. Then, for a pair (s, s') of the original sentence and new generated sentence ($s' = prompt \cup W'$), we can define a diversity score $d_{s,s'} = WER(W, W')$. Given the definition of d, for a language l, we can define a diverse batch

$$B^l_{diverse} = \bigcup_{i \in [1, |\mathcal{H}_l|]} \{s'_i \mid d_{s_i,s'_i} \geq t\}$$

where t is an hyper-parameter, representing an error threshold. t can be tuned because a small t will result in a less diverse batch, while a very huge value will result in an empty or almost empty batch.

7 Ethics Statement

As any modern technology, machine learning algorithms, are subject to potential dual good or bad usage. Our work is motivated by the desire of making AI (in general, NLP in particular) applications to be inclusive to the low-resourced languages (which are the vast majority of existing living languages), hence benefiting to humanity and society. We strongly discourage bad and unethical use of our work (and its derivations).

References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021a. MasakhaNER: Named Entity Recognition for African Languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

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

David Ifeoluwa Adelani, Jesujoba Oluwadara 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 Chinenye Emezue, Colin Leong, Michael Beukman, Shamsuddeen Hassan Muhammad, Guyo Dub Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles HACHEME, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ayoade Ajibade, Oluwaseyi Ajayi Ajayi, Yvonne Wambui Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Koffi KALIPE, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, and Ayodele Awokoya. 2022a. A few thousand translations go a long way! leveraging pre-trained models for african news translation. In *NAACL-HLT*.

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.

Jesujoba O Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022a. Multilingual language model adaptive fine-tuning: A study on african languages. *arXiv preprint arXiv:2204.06487*.

Jesujoba O Alabi, Adelani David Ifeoluwa, Mosbach Marius, and Klakow Dietrich. 2022b. Multilingual Language Model Adaptive Fine-Tuning: A case study on African Languages. *COLING*.

Nzeyimana Antoine and Rubungo Andre Niyongabo. 2022. KinyaBERT:a Morphology-aware Kinyarwanda Language Model. *ACL*.

Hounkpati B. C. Capo. 1991. *A comparative phonology of Gbe*. Foris Publications.

Yukun Chen, Thomas A. Lasko, Qiaozhu Mei, Joshua C. Denny, and Hua Xu. 2015. A study of active learning methods for named entity recognition in clinical text. *Journal of Biomedical Informatics*, 58:11–18.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig (eds.). 2020. Ethnologue: Languages of the world. twenty-third edition.

Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. Active Learning for BERT: An Empirical Study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7949–7962, Online. Association for Computational Linguistics.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1183–1192. JMLR.org.

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.

Moksh Jain, Emmanuel Bengio, Alex-Hernandez Garcia, Jarrid Rector-Brooks, Bonaventure F. P. Dossou, Chanakya Ekbote, Jie Fu, Tianyu Zhang, Micheal Kilgour, Dinghuai Zhang, Lena Simine, Payel Das, and Yoshua Bengio. 2022. Biological sequence design with gflownets.

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.

Ogueji Kelechi, Zhu Yuxin, and Lin Jimmy. 2021. Small Data? No Problem! Exploring the Viability of Pretrained Multilingual Language Models for Low-resourced Languages. *EMNLP*, pages 116–126.

Claire Lefebvre and Anne-Marie Brousseau. 2002. *A grammar of Fongbe*. Mouton de Gruyter.

Nikolay Malkin, Zhen Wang, and Nebojsa Jojic. 2022. Coherence boosting: When your pretrained language model is not paying enough attention. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8214–8236, Dublin, Ireland. Association for Computational Linguistics.

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.

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.

Iyanuoluwa Shode, David Ifeoluwa Adelani, and Anna Feldman. 2022. YOSM: A NEW YORUBA SENTIMENT CORPUS FOR MOVIE REVIEWS. In *3rd Workshop on African Natural Language Processing*.

Aditya Siddhant and Zachary C. Lipton. 2018. Deep bayesian active learning for natural language processing: Results of a large-scale empirical study.

Justin S. Smith, Benjamin Tyler Nebgen, Nick Lubbers, Olexandr Isayev, and Adrian E. Roitberg. 2018. Less is more: sampling chemical space with active learning. *The Journal of chemical physics*, 148 24:241733.

Manuel Tonneau, Dhaval Adjodah, Joao Palotti, Nir Grinberg, and Samuel Fraiberger. 2022. Multilingual detection of personal employment status on Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6564–6587, Dublin, Ireland. Association for Computational Linguistics.

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

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2019. Ccnet: Extracting high quality monolingual datasets from web crawl data.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2019. Huggingface's transformers: State-of-the-art natural language processing.

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 *NAACL*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*.

A Language Characteristics

Amharic (amh) also called Amarinya or Amerigna, is a Semitic language, an official language of Ethiopia, and is also spoken in Eritrea. Amharic is written with a modified version of the Ge'ez script, known as Fidel, consisting of 33 basic characters, each of them with at least 7 vowel sequences. Unlike Central and Northwest Semitic languages such as Arabic, Hebrew and Assyrian Aramaic, Amharic is written from left to right. The language has a variety of local dialects, all of which are mutually intelligible. There are three major dialects: Gondar, Gojjami, and Showa. There are specially marked differences in pronunciation, vocabulary, and grammar between the northern Gojjami and the southern Showa dialects.

Afan Oromo (oro) is an Afroasiatic language that belongs to the Cushitic branch spoken by about 30 million people in Ethiopia, Kenya, Somalia and Egypt, and it is the third largest language in Africa. The Oromo people are the largest ethnic group in Ethiopia and account for more than 40% of the population. They can be found all over Ethiopia, and particularly in Wollega, Shoa, Illubabour, Jimma, Arsi, Bale, Hararghe, Wollo, Borana and the southwestern part of Gojjam[2]. Afan Oromo is written with a Latin alphabet called Qubee. Like most other Ethiopian languages, whether Semitic, Cushitic, or Omotic, Oromo has a set of ejective consonants, that is, voiceless stops or affricates that are accompanied by glottalization and an explosive burst of air. Afan Oromo has another glottalized phone that is more unusual, an implosive retroflex stop, "dh" in Oromo orthography, a sound that is like an English "d" produced with the tongue curled back slightly and with the air drawn in so that a glottal stop is heard before the following vowel begins. It is retroflex in most dialects, though it is not strongly implosive and may reduce to a flap between vowels[3]. In the Qubee alphabet, letters include the digraphs ch, dh, ny, ph, sh. Gemination is not obligatorily marked for digraphs, though some writers indicate it by doubling the first element: qopphaa'uu 'be prepared'. Afan Oromo has five vowel phonemes, i.e., sounds that can differentiate word meaning. They can be short or long. The length of the vowel makes a difference in word meaning e.g., laga 'river' and laagaa 'roof of the

[2] https://omniglot.com/writing/oromo.htm

[3] https://en.wikipedia.org/wiki/Oromo_language

mouth'. Afan Oromo has 25 consonant phonemes, i.e., sounds that make a difference in word meaning. Like its close relative, Somali, native Oromo words do not have the consonants /p/, /v/, and /z/.

Bambara (bam) is a Western Mande language with about 14 million speakers mainly in Mali, and also in Senegal, Niger, Mauritania, Gambia and Côte d'Ivoire. It is spoken principally among the Bambara ethnic group in Mali, where it is the national language and the most widely understood one. Bambara is usually written with the Latin alphabet, though the N'Ko and Arabic alphabets are also used to some extent. It uses seven vowels a, e, ɛ, i, o, ɔ, and u each of which can be nasalized, pharyngealized and murmured, giving a total number of 21 vowels.

Ghomalá' (bbj) is a major Bamileke language of spoken in Cameroon. It is spoken by an estimated 1.1 million people in two main population groups.

Éwé (ewe) is a language spoken in Togo and southeastern Ghana by approximately 20 million people mainly in West Africa in the countries of Ghana, Togo, and Benin. It is recognised as a national language in Ghana, where English is the official language, and in Togo, where French is the official language. 'Ewe' is also the name of the tribal group that speaks this language. Éwé has three distinguishable dialects. Most of the differences among the dialects have to do with phonology. All dialects are mutually intelligible. Éwé is written in the African reference alphabet, first proposed by a UNESCO-organized conference in 1978. It is a version of the Latin alphabet adapted to represent Éwé sounds. Some sounds are represented by two-letter sequences, e.g., dz, ts, gb, kp, ny. Éwé has seven oral and five nasal vowels. Nasal vowels are produced by lowering the soft palate so that air escapes both through the mouth and the nose. Nasal vowels are marked by a tilde.

Fon (fon) also known as Fongbé is a native language of Benin Republic. It is spoken in average by 1.7 million people. Fon belongs to the *Niger-Congo-Gbe* languages family. It is a tonal, isolating and left-behind language according to (Joshi et al., 2020), with an *Subject-Verb-Object* (SVO) word order. Fon has about 53 different dialects, spoken throughout Benin (Lefebvre and Brousseau, 2002; Capo, 1991; Eberhard et al., 2020). Its alphabet is based on the Latin alphabet, with the addition of the letters: ɔ, ɗ, ɛ, and the digraphs gb, hw, kp, ny, and xw. There are 10 vowels phonemes in Fon: 6 said to be closed [i, u, ĩ, ũ], and 4 said to be opened [ɛ, ɔ, a, ã]. There are 22 consonants (m, b, n, ɗ, p, t, d, c, j, k, g, kp, gb, f, v, s, z, x, h, xw, hw, w). Fon has two phonemic tones: high and low. High is realized as rising *(low–high)* after a consonant. Basic disyllabic words have all four possibilities: *high-high*, *high-low*, *low-high*, and *low-low*. In longer phonological words, like verb and noun phrases, a high tone tends to persist until the final syllable. If that syllable has a phonemic low tone, it becomes falling *(high–low)*. Low tones disappear between high tones, but their effect remains as a downstep. Rising tones *(low–high)* simplify to high after high (without triggering downstep) and to low before high (Lefebvre and Brousseau, 2002; Capo, 1991).

Hausa (hau) belongs to the West Chadic branch of the Afro-Asiatic language family. It is one of the largest languages on the African continent, spoken as a first language by the original Hausa people and by people of Fula ancestry. Hausa is the majority language of much of northern Nigeria and the neighboring Republic of Niger. In addition, there is a sizable Hausa-speaking community in Sudan[4]. It has an alphabet of 29 letters containing 5 vowels and 24 consonants. Hausa alphabet is a Latin script/Roman alphabet/English letters except (x, v, p, and q) and also added six extra letters (ɓ, ɗ, ƙ, sh, ts and ƴ (Adelani et al., 2021b). Hausa is an agglutinative language, i.e., it adds suffixes to roots for expressing grammatical relations without fusing them into one unit, as is the case in Indo-European languages.

Ìgbò (ibo) is one of the largest languages of West Africa, is spoken by 18 million people in Nigeria. It belongs to the Benue-Congo group of the Niger-Congo language family. The language is thought to have originated around the 9th century AD in the area near the confluence of the Niger and Benue rivers, and then spread over a wide area of south-eastern Nigeria [5]. Igbo is a national language of Nigeria and is also recognised in Equatorial Guinea. Igbo is written in an expanded version of the Latin alphabet. Igbo is made up of many different dialects which aren't mutually intelligible to other Igbo speakers at times.

[4] https://www.mustgo.com/worldlanguages/hausa/

[5] https://www.mustgo.com/worldlanguages/igbo/

Kinyarwanda (kin) a is part of the Bantu subgroup of the central branch of the Niger-Congo language family. It is closely related to Kirundi, the language of Burundi. The Rwanda language is mutually intelligible with Kirundi, which is spoken in neighboring Burundi[6]. It has only 18/19 consonants, as X and Q are not found in the alphabet. L is often replaced by R, but due to the appearance of imported words in the language, that is not always the case. It has five vowel phonemes, i.e., sounds that make a difference in word meaning.

Lingala (lin) is a Central Bantu language that belongs to the largest African languages phylum: the Niger-Congo. Lingala is spoken as a first, second, and third language primarily in the Democratic Republic of Congo (DRC), the Republic of Congo (Congo-Brazzaville), and in parts of five neighboring central African states: Northwestern Angola, eastern Gabon, southern Central African Republic, and southwestern Sudan. The estimated number of speakers ranges from twenty to twenty five million[7]. It is written with the Latin alphabet. The seven vowels are represented by five symbols. The orthographic symbols 'e' and 'o' each represent two sounds. There are two tones in Lingala. High tone is represented with an acute accent, while low tone is unmarked.

Luganda (lug) is a Bantu language spoken in the African Great Lakes region. It is one of the major languages in Uganda and is spoken by more than 10 million Baganda and other people principally in central Uganda including the capital Kampala of Uganda. Its alphabet is composed of twenty-four letters; 18 consonants (b, p, v, f, m, d, t, l, r, n, z, s, j, c, g, k, ny, ŋ), 5 vowels (a, e, i, o, u) and 2 semi-vowels(w, y). Since the last consonant ŋ) does not appear on standard typewriters or computer keyboards, it is often replaced by the combination ng'. All consonants are pronounced as if with letter 'a' or 'ah' at the end. For example, bah, cah, jah, gah, kah, mah, pah, lah, zah, e.t.c

Luo (luo) are spoken by the Luo peoples in an area ranging from southern Sudan to southern Kenya, with Dholuo extending into northern Tanzania and Alur into the Democratic Republic of the Congo. Luo has a CVC or VC structure—consonant/vowel/consonant or vowel/consonant. This means that Luo words can end in a consonant, like gin, they are. This is unlike Bantu languages, where words must end in a vowel. Luo language is, therefore, more similar to English articulation, while Bantu languages are more like Italian[8].

Mooré (mos) is a Gur language of the Oti–Volta branch and one of two official regional languages of Burkina Faso. It is the language of the Mossi people, spoken by approximately 8 million people in Burkina Faso, plus another 1M+ in surrounding countries such as Ghana, Cote D'ivoire, Niger, Mali and Togo as a native language, but with many more L2 speakers. Mooré is spoken as a first or second language by over 50% of the Burkinabè population.

Chewa (nya) is a Bantu language spoken in much of Southern, Southeast and East Africa, namely the countries of Malawi and Zambia, where it is an official language, and Mozambique and Zimbabwe where it is a recognised minority language. Chewa has five vowel sounds: /a, ɛ, i, ɔ, u/; these are written a, e, i, o, u.

Naija (pcm) is an English-based creole language spoken as a lingua franca across Nigeria. The language is sometimes referred to as "Pijin" or Broken (pronounced "Brokun").

Shona (sna) is a Bantu language of the Shona people of Zimbabwe. All syllables in Shona end in a vowel. Consonants belong to the next syllable. For example, mangwanani ("morning") is syllabified as ma.ngwa.na.ni; "Zimbabwe" is zi.mba.bwe. No silent letters are used in Shona.

Swahili (swa) also known by its native name Kiswahili, is a Bantu language and the native language of the Swahili people native primarily to Tanzania. Swahili has become a second language spoken by tens of millions in four African Great Lakes countries (Kenya, DRC, Uganda, and Tanzania), where it is an official or national language, while being the first language for many people in Tanzania especially in the coastal regions of Tanga, Pwani, Dar es Salaam, Mtwara and Lindi. Standard Swahili has five vowel phonemes: /a/, /ɛ/, /i/, /ɔ/, and /u/.

Setswana (tsn) is a Bantu language spoken in Southern Africa by about 14 million people.

[6] https://nalrc.indiana.edu/doc/brochures/kinyarwanda.pdf

[7] https://nalrc.indiana.edu/doc/brochures/lingala.pdf

[8] https://owlcation.com/humanities/Luo-language-of-Kenya-Conversation-Basics

Setswana is an official language and lingua franca of Botswana and South Africa.

Akan/Twi is a dialect of the Akan language spoken in southern and central Ghana by several million people, mainly of the Akan people, the largest of the seventeen major ethnic groups in Ghana. Twi excludes consonants such as c, j, q, v, x and z. It has 15 consonants and 7 vowels. Apart from [a], [e], [i], [o] and [u], Twi also has 2 additional vowels; [ɛ] and [ɔ].

Wolof (wol) is a language of Senegal, Mauritania, and the Gambia, and the native language of the Wolof people. Wolof is the most widely spoken language in Senegal, spoken natively by the Wolof people (40% of the population) but also by most other Senegalese as a second language.

Xhosa (xho) also isiXhosa as an endonym, is a Nguni language and one of the official languages of South Africa and Zimbabwe. The Xhosa language employs 26 letters from the Latin alphabet. Xhosa has an inventory of ten vowels: [a], [ɛ e], [i], [ɔ o] and [u] written a, e, i, o and u in order, all occurring in both long and short. The /i/ vowel will be long in the penultimate syllable and short in the last syllable.

Yorùbá (yor) has 25 Latin letters without the Latin characters (c, q, v, x and z) and with additional letters (ẹ, gb,ṣ, ọ).Yorùbá is a tonal language with three tones: low ("\"), middle ("—", optional) and high ("/"). The Latin letters ⟨c⟩, ⟨q⟩, ⟨v⟩, ⟨x⟩, ⟨z⟩ are not used as part of the official orthography of Standard Yorùbá, however, they exist in several Yorùbá dialects. The tonal marks and underdots are referred to as diacritics and they are needed for the correct pronunciation of a word. Yorùbá is a highly isolating language and the sentence structure follows subject-verb-object (Adelani et al., 2021b).

Zulu (zul) is the mother tongue of the Zulu people, South's Africa largest ethnic group, who created an empire in the 19th century.Zulu has a 7-vowel system. Each vowel can be long or short. Zulu has close to 50 consonants including clicks, ejectives and implosives. Clicks originated in Khoisan languages and then spread into some neighboring Bantu ones. In Zulu they have three places of articulation: central alveolar, lateral alveolar and palatal combined with five accompaniments (plain, aspirated, voiced, nasal, and voiced nasal).