Are existing pre-trained LLMs really effective on African languages?

SANGKAK-CHALLENGE-IA

MBOUOPDA Michael Franklin - ML Researcher at Huawei
30 Sept 2023

Agenda

Introduction

Proposal

Results

Demo

Introduction

Challenge's description

POS Tagging

- NLP for African languages is still a difficult task
- *Masakhane NLP* group fine-tuned LLM to achieve POS tagging for 20 African languages
- Can we do better?

MasakhaPOS: Part-of-Speech Tagging for Typologically Diverse African Languages

Cheikh M. Bamba Dione[1,†,*], David Ifeoluwa Adelani[2,†,*], Peter Nabende[3,†], Jesujoba O. Alabi[4,†], Thapelo Sindane[5], Happy Buzaaba[6†], Shamsuddeen Hassan Muhammad[7,8†], Chris Chinenye Emezue[9,10†], Perez Ogayo[11†], Anuoluwapo Aremu[†], Catherine Gitau[†], Derguene Mbaye[12†], Jonathan Mukiibi[3†], Blessing Sibanda[†], Bonaventure F. P. Dossou[10,13,14†], Andiswa Bukula[15], Rooweither Mabuya[15], Allahsera Auguste Tapo[16†], Edwin Munkoh-Buabeng[17†], Victoire Memdjokam Koagne[†], Fatoumata Ouoba Kabore[18†], Amelia Taylor[19], Godson Kalipe[†], Tebogo Macucwa[5], Vukosi Marivate[5,13†], Tajuddeen Gwadabe[†], Elvis Tchiaze Mboning[†], Ikechukwu Onyenwe[20], Gratien Atindogbe[21], Tolulope Anu Adelani[†], Idris Akinade[22], Olanrewaju Samuel[†], Marien Nahimana, Théogène Musabeyezu, Emile Niyomutabazi, Ester Chimhenga, Kudzai Gotosa, Patrick Mizha, Apelete Agbolo[23], Seydou Traore[24], Chinedu Uchechukwu[20], Aliyu Yusuf[8], Muhammad Abdullahi[8], Dietrich Klakow[4]

[†]Masakhane NLP, [1]Université Gaston Berger, Senegal, [2]University College London, UK, [3]Makerere University, Uganda, [4]Saarland University, Germany, [5]University of Pretoria, South Africa, [6] RIKEN Center for AIP, Japan, [7]Bayero University Kano, Nigeria. [8]University of Porto, Portugal, [9]Technical University of Munich, Germany, [10]Lanfrica, [11]Carnegie Mellon University, USA, [12]Baamtu, Senegal, [13]Lelapa AI, [14]Mila Quebec AI Institute, Canada, [15]SADiLaR, South Africa, [16]Rochester Institute of Technology, USA, [17]TU Clausthal, Germany, [18]Uppsala University, Sweden, [19]Malawi University of Business and Applied Science, Malawi, [20]Nnamdi Azikiwe University, Nigeria, [21]University of Buea, Cameroon, [22]University of Ibadan, Nigeria, [23]Ewegbe Akademi, Togo, [24]AMALAN, Mali.

Proposal

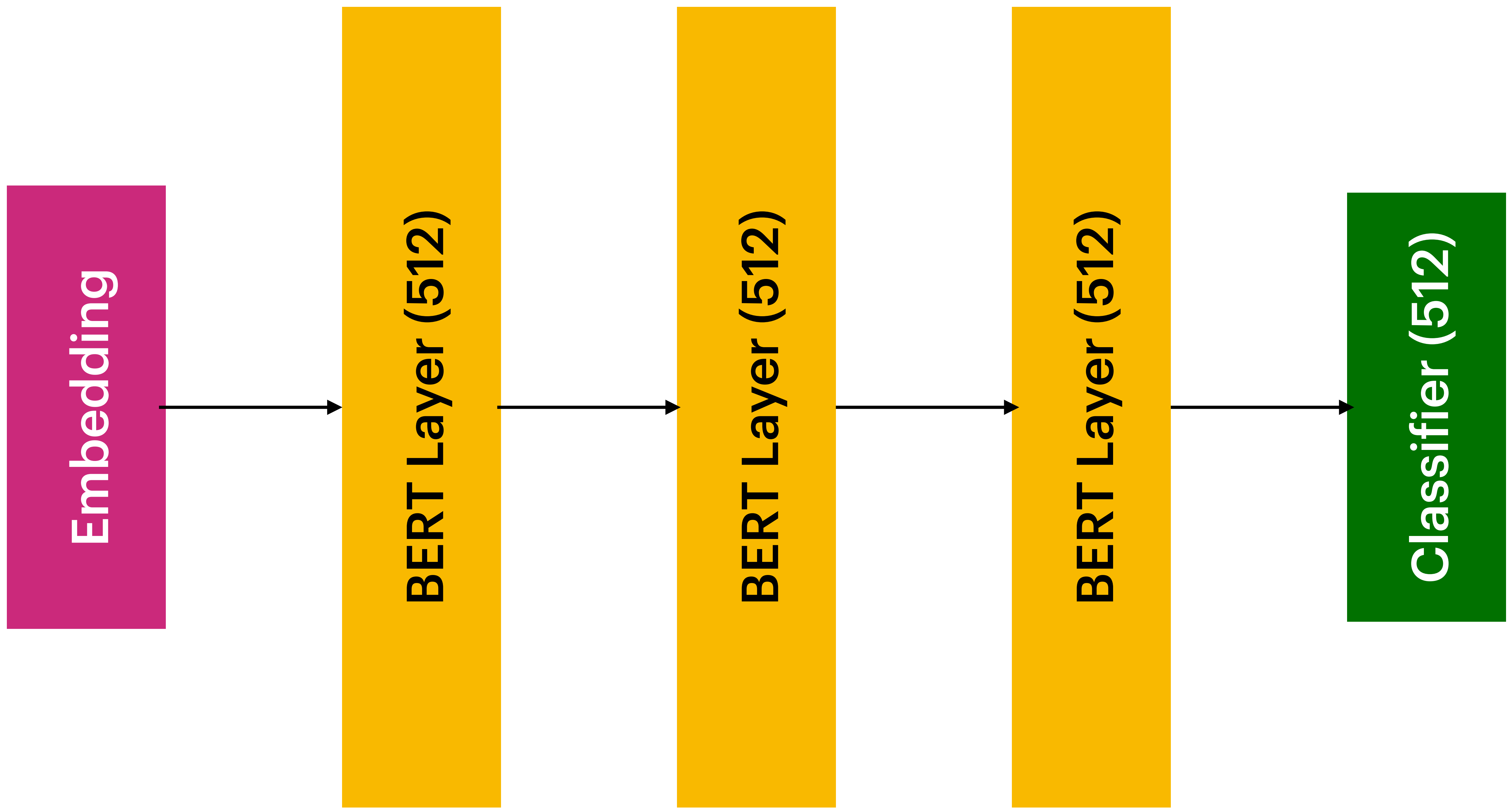
Proposal

Check if 550M parameters are really needed for the task

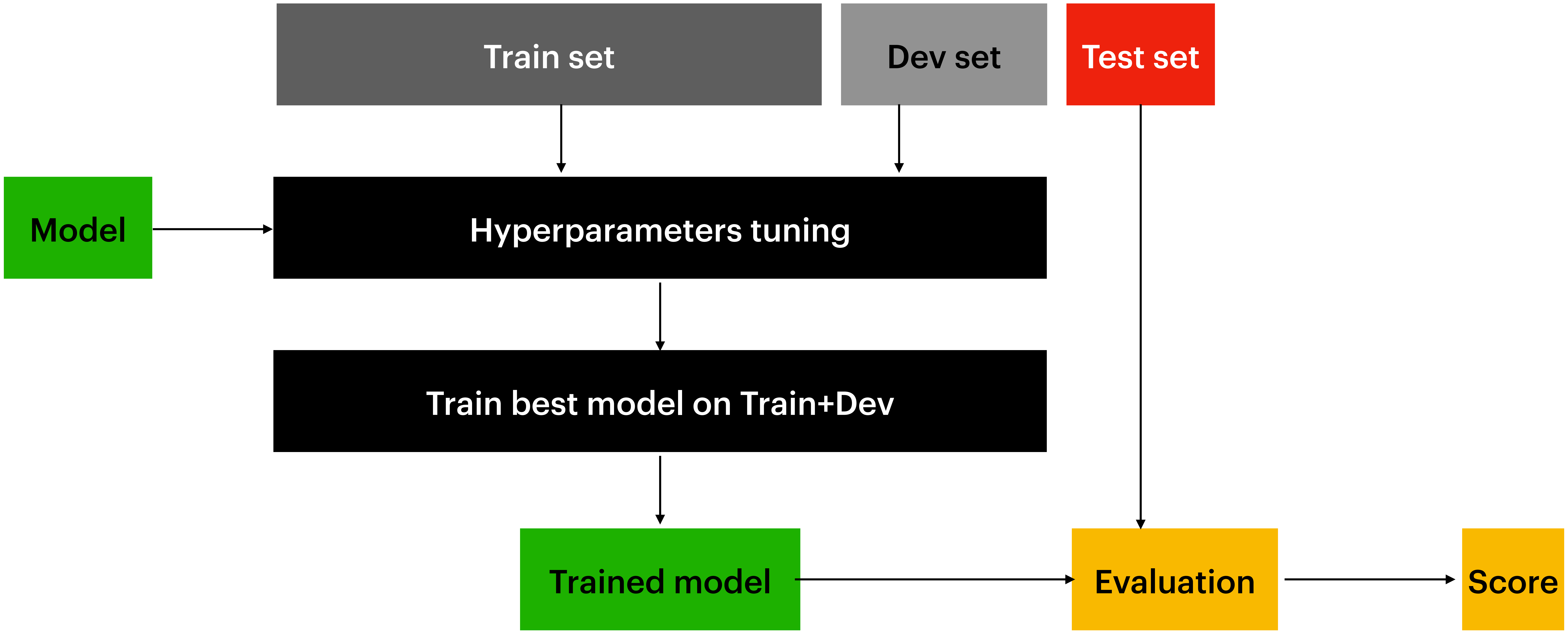
- Why
  - LLM are very expensive (fine-tuning, inference)
  - The dataset is quite small
  - LLM are pre-trained on tasks that are much more complex than POS tagging
- How
  - Solve the problem using smaller models trained from scratch

Model's architecture

- BERT architecture
- 4 attention heads
- Dropout
  - Attention: 0.1
  - BERT Layer: 0.2
  - Classifier: 0.1
- 16 354 605 parameters



Training and evaluation scheme

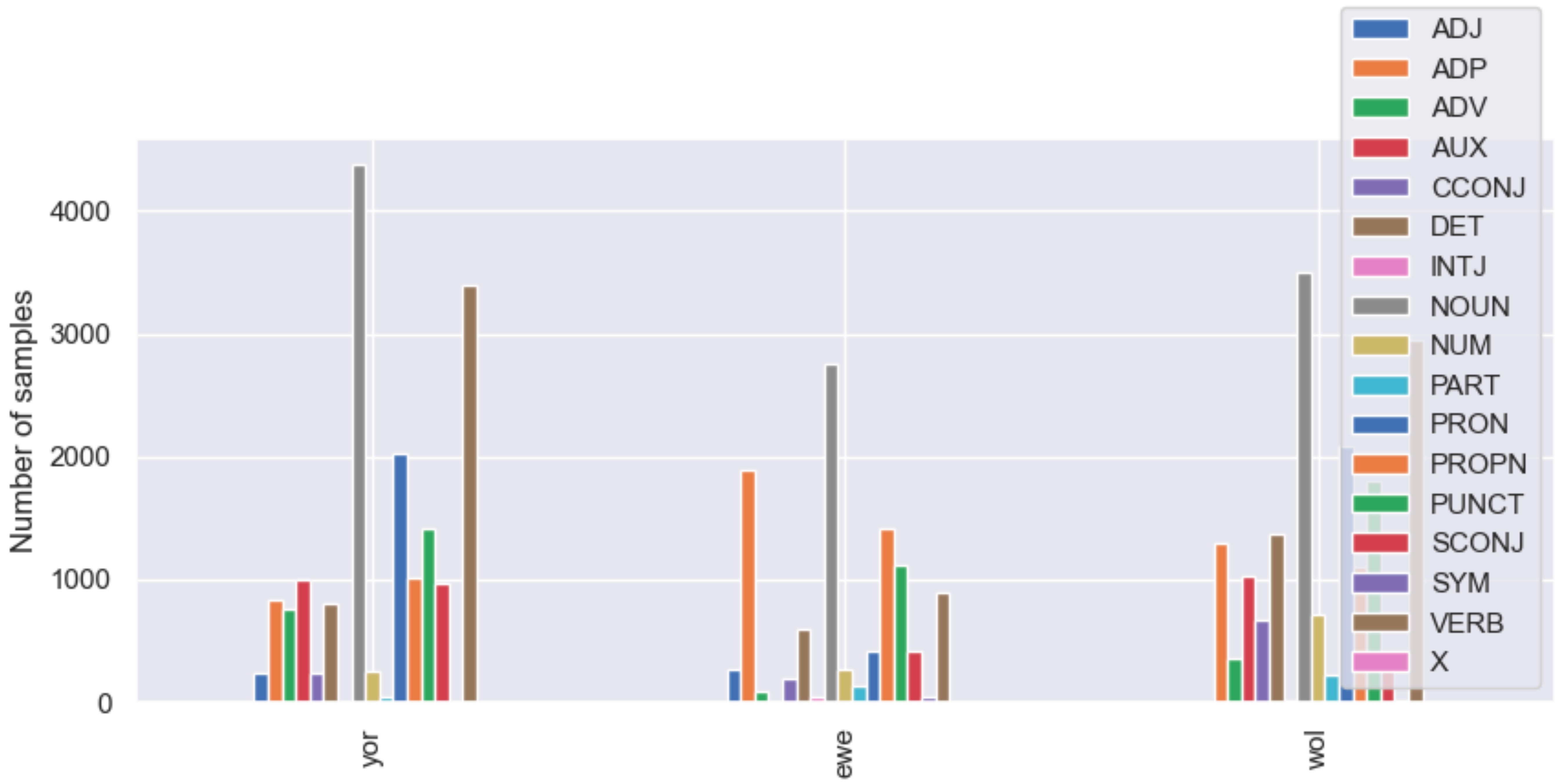


Results

Metrics

- The dataset is unbalanced
- Accuracy is not a good metric
- We report: Precision, Recall and F1-score in addition to Accuracy

Number of samples per tag for three languages



Refer to Dione et al. (2023) for the description of the dataset

Masakhane reported the accuracy only

Accuracy

Mono-lingual models

	bam	bbj	ewe	fon	hau	ibo	kin	lug	mos	nya	pcm	sna	swa	twi	wol	xho	yor	zul	mean	std
f1-score	0.86	0.68	0.83	0.82	0.82	0.80	0.82	0.75	0.83	0.62	0.8	0.75	0.8	0.75	0.85	0.60	0.83	0.64	0.77	0.08
precision	0.86	0.69	0.83	0.83	0.83	0.81	0.84	0.75	0.83	0.63	0.8	0.76	0.8	0.76	0.85	0.63	0.83	0.67	0.78	0.08
recall	0.86	0.68	0.83	0.82	0.83	0.80	0.83	0.76	0.83	0.63	0.8	0.76	0.8	0.76	0.85	0.61	0.83	0.66	0.77	0.08
accuracy	0.86	0.68	0.83	0.82	0.83	0.80	0.83	0.76	0.83	0.63	0.8	0.76	0.8	0.76	0.85	0.61	0.83	0.66	0.77	0.08

Multi-lingual model

	bam	bbj	ewe	fon	hau	ibo	kin	lug	mos	nya	pcm	sna	swa	twi	wol	xho	yor	zul	mean	std
f1-score	0.84	0.68	0.81	0.82	0.81	0.74	0.87	0.77	0.8	0.70	0.79	0.79	0.81	0.74	0.83	0.69	0.81	0.73	0.78	0.06
precision	0.84	0.68	0.82	0.82	0.81	0.77	0.87	0.77	0.8	0.69	0.79	0.79	0.82	0.74	0.84	0.69	0.81	0.74	0.78	0.05
recall	0.84	0.67	0.81	0.82	0.81	0.74	0.87	0.77	0.8	0.70	0.79	0.80	0.82	0.74	0.84	0.69	0.81	0.74	0.78	0.06
accuracy	0.84	0.67	0.81	0.82	0.81	0.74	0.87	0.77	0.8	0.70	0.79	0.80	0.82	0.74	0.84	0.69	0.81	0.74	0.78	0.06

Comparison to Masakhane

Assuming that Masakhane followed the same training and evaluation scheme

	bam	bbj	ewe	fon	hau	ibo	kin	lug	mos	nya	pcm	sna	swa	twi	wol	xho	yor	zul	mean	std
AfroXLMR-large (550M)	0.90	0.85	0.89	0.90	0.93	0.79	0.98	0.92	0.91	0.83	0.91	0.90	0.93	0.85	0.93	0.89	0.95	0.90	0.90	0.05
Ours - monolingual (16M)	0.86	0.68	0.83	0.82	0.83	0.80	0.83	0.76	0.83	0.63	0.80	0.76	0.80	0.76	0.85	0.61	0.83	0.66	0.77	0.08
Ours - multilingual (16M)	0.84	0.67	0.81	0.82	0.81	0.74	0.87	0.77	0.80	0.70	0.79	0.80	0.82	0.74	0.84	0.69	0.81	0.74	0.78	0.06

34x less parameters

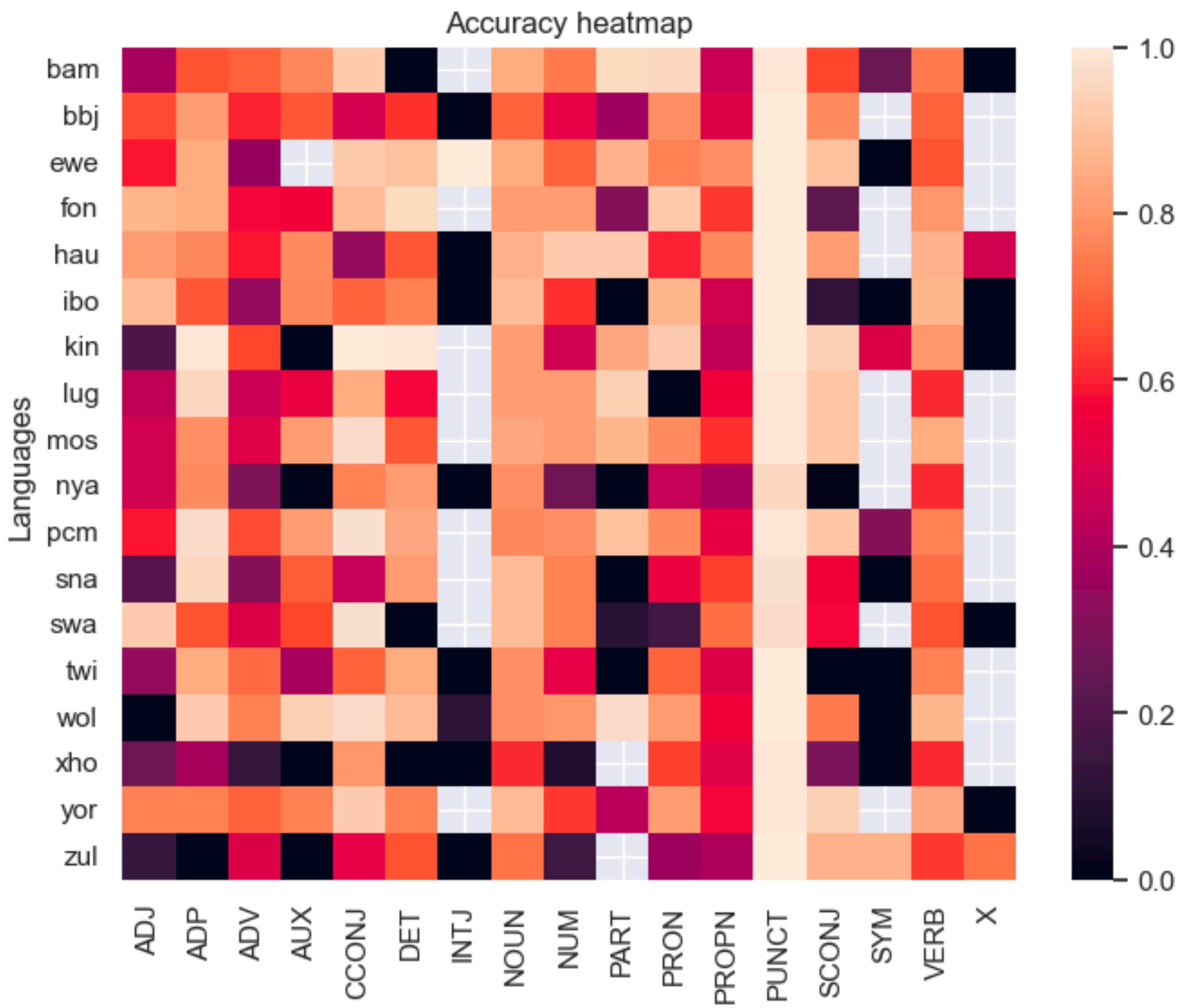
A lost of only 0.12 in accuracy

Vs

AfroXLM
550M parameters

What is good? What need improvement?

Monolingual



Multilingual

Demo