It is *possible* to obtain deep contextualised representations of *new low-resourced language classification data* without having to change the Pre-trained Language Model architecture, through *pre- and post-processing techniques*

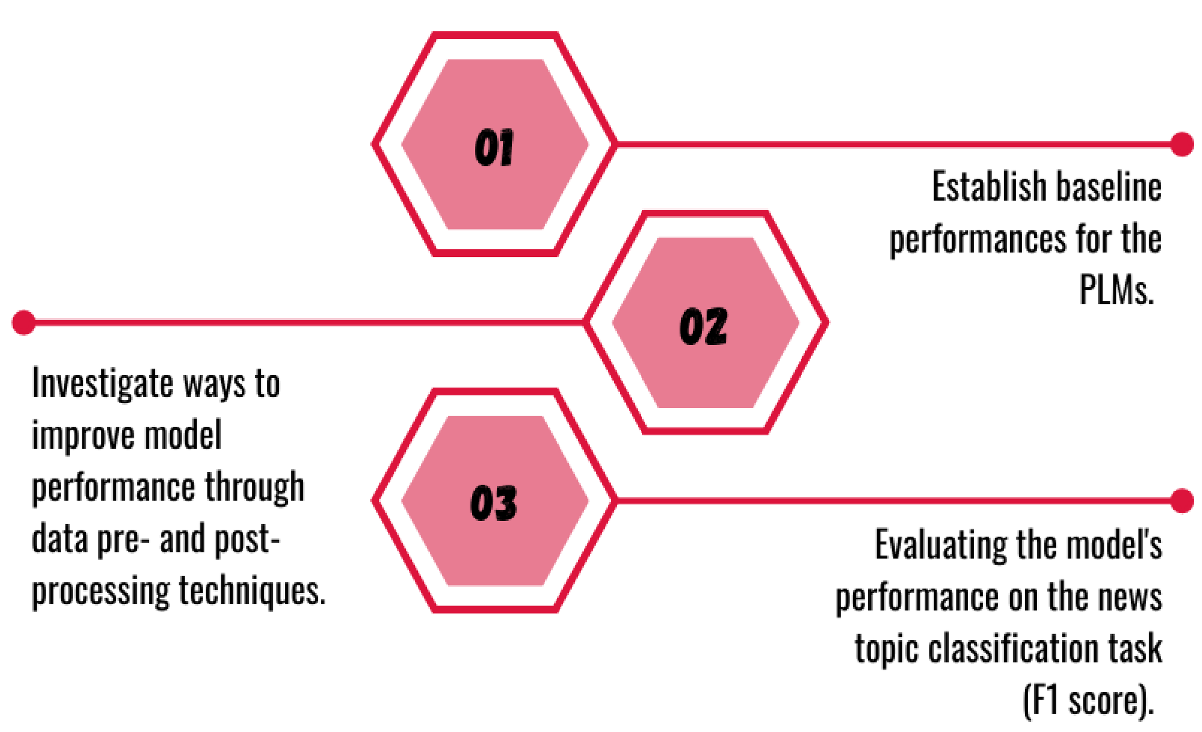
Increasing Linguistic Diversity in the NLP space: Fine-tuning Multilingual Pretrained African Language Models

INTRO

- Today's NLP research mainly focuses on *20 of the 7,000* languages spoken worldwide, leaving many African languages unstudied. These are often referred to as low-resource languages(LRLs).
- With the recent increase in low-resource African language text corpora, there have been advancements which have led to development of multilingual pre-trained language models (PLMs), based on African languages.
- This led to the question we attempting to answer: **Can these PLMs be fine-tuned to perform similarly well on different LRLs (e.g., South African Language)?**

METHODS

To solve this problem, we ran an experiment on a set of 3 PLMs *(AfriBERTa, AfroXLMR and AfroLM), using an isiZulu news classification dataset.*



EXPERIMENTAL SETUP

Evaluated the *F1 scores* for *news topic classification*

- Modification of Architecture: freezing the encoder layers, except for the classifier layer.
- Modification of data: Due to the uneven distribution of the dataset – we explored using a subset of the data which contain samples from the top 10 classes. On top of this, increasing the number of samples by using text augmentation.
- Combination of datasets: The African News Topic dataset was used in addition with the original dataset.

RESULTS

Implemented Method	Dataset	AfriBERTa	AfroXLMR	AfroLM
Freezing lower layers	isiZulu News	0.424	**0.464**	0.384
Using the data in top 10 classes	isiZulu News	0.514	**0.771**	0.643
Augmented Data	isiZulu News	0.538	**0.724**	0.591
Sampling method	Combined Data	0.574	**0.703**	0.534
Augmented Data	Combined Data	0.744	N/A	N/A
Using the data in top 10 classes	Combined Data	0.745	N/A	N/A

Table 3: **Experimentation performance (F1 Score) on the isiZulu News Dataset. The best scores obtained during the experimentation is in bold. The experimentation configuration which results in an increase in the baseline results for all the models is highlighted in grey.**

DISCUSSION

- Based on the baseline performance, the models performed best in the classification tasks on languages they were originally trained on.
- The performances of all the PLMs increased with data modification , which suggests the importance of data pre and post processing.

AfriBERTa
Afaan Oromoo, Amharic, Gahuza, Hausa, Igbo, **Nigerian Pidgin,** **Somali**, Swahili, Tigrinya and Yorùbá

Afro-XLMR
Afrikaans, Amharic, Hausa, Igbo, **Malagasy**, Chichewa, Sesotho, Oromo, **Nigerian-Pidgin,** Kinyarwanda, Kirundi, Shona, **Somali,** Swahili, isiXhosa, Yoruba, **isiZulu,** Arabic, French, and English

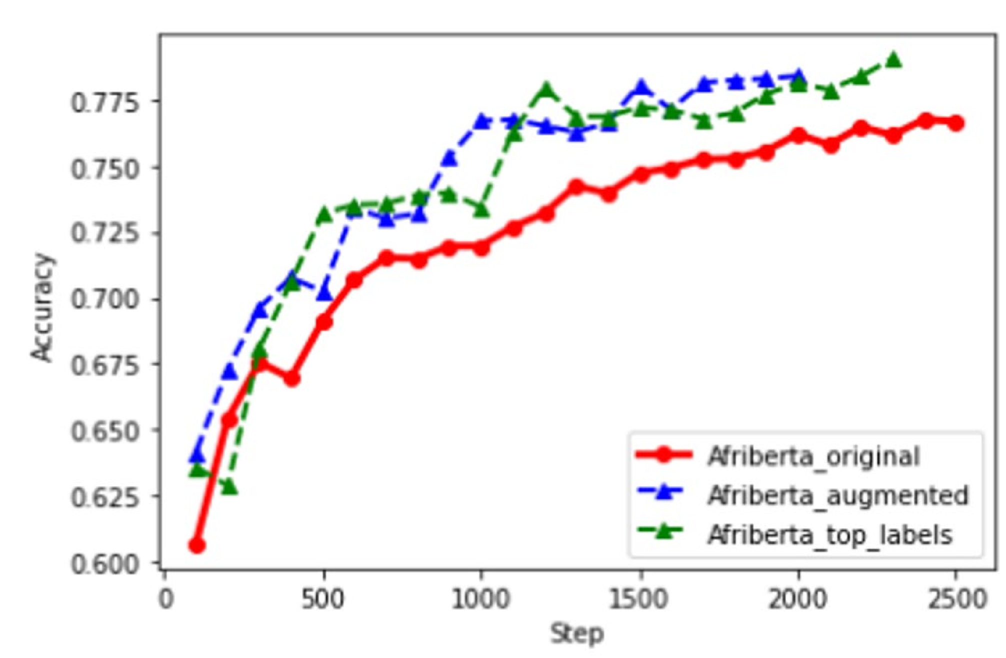
AfroLM
Amharic, Afan Oromo, Bambara, Ghomalá, Éwé, Fon, Hausa, Igbo Kinyarwanda, **Lingala,** Luganda, Luo, Mooré, Chewa, Naija, Shona, Swahili, Setswana, Twi, Wolof, Xhosa, Yorùbá, and **Zulu**

Languages
isiZulu
Lingala
Somali
Nigerian Pidgin
Malagasy

Class Labels
crime, law and justice
politics
society
economy, business and finance
education
health
disaster, accident and emergency incident
arts, culture, entertainment and media
religion and belief
human interest
sport
labour
conflict, war and peace
weather
environment
lifestyle and leisure
Africa
Automotive
Science and technology
World

Dataset	Language	AfriBERTa	AfroXLMR	AfroLM	Vocab Size
isiZulu News	isiZulu	0.506	**0.695**	0.616	250 002
ANTC	isiZulu	0.825	**0.854**	0.832	70 006
ANTC	Lingala	0.618	0.607	**0.664**	70 006
ANTC	Malagasy	0.524	**0.682**	0.512	70 006
ANTC	Pidgin	0.829	0.835	**0.838**	70 006
ANTC	Somali	**0.780**	0.682	0.751	70 006
Combined Data	All languages	**0.726**	0.704	0.552	N/A
avg		0.694	**0.723**	0.681	
avg (excl. combined data)		0.680	**0.726**	0.702	

Table 2: **Baseline performance (F1 Score). The best performance for each language and overall the best performing model is in bold.**



Fiskani Banda, *Rozina Myoya,Vukosi Marivate, Abiodun Modupe*

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Data Science for Social Impact

Scan me