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.

01 Establish baseline performances for the PLMs.
02 Investigate ways to improve model performance through data pre- and post-processing techniques.
03 Evaluating the models performance on the news topic classification task (F1 scores).

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

<table>
<thead>
<tr>
<th>Implemented Method</th>
<th>Dataset</th>
<th>AfriBERTa</th>
<th>AfroXLMR</th>
<th>AfroLM</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1-score</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Using the data in top 10 classes</td>
<td>isiZulu News</td>
<td>0.514</td>
<td>0.771</td>
<td>0.641</td>
</tr>
<tr>
<td>Augmented Data</td>
<td>isiZulu News</td>
<td>0.536</td>
<td>0.724</td>
<td>0.591</td>
</tr>
<tr>
<td>Sampling method</td>
<td>Combined Data</td>
<td>0.703</td>
<td>0.514</td>
<td></td>
</tr>
<tr>
<td>Using the data in top 10 classes</td>
<td>Combined Data</td>
<td>0.745</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

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.

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.

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