FINE-TUNING MULTILINGUAL PRETRAINED AFRICAN LANGUAGE MODELS

Rozina Myoya[1], Fiskani Banda[1], Vukosi Marivate[*,1,2], Abiodun Modupe[1]
[1] Department of Computer Science, University of Pretoria
[2] Lelapa AI
[*] Corresponding email vukosi.marivate@cs.up.ac.za

1 INTRODUCTION

In the 1990s, the tools of natural language processing (NLP) underwent a big change, moving from rule-based to statistical-based methods to make computers understand language better (Magueresse et al., 2020). Today's NLP research mainly focuses on 20 of the 7,000 languages spoken worldwide, which account for more than 95% of the world's population, leaving the vast majority of African languages unstudied. These languages are often called low-resource languages (LRLs), even though this term is not always clear. To improve language translation in LRLs, NLP research has shifted focus to neural machine translation (NMT) to handle errors such as phoneme substitutions, grammatical structure, and sentence boundaries, all of which pose challenges to NMT robustness (Li et al., 2021). NMT systems have been designed to learn from the language data available in LRLs, creating models that can generalize and handle errors efficiently.

Despite tremendous spurts of growth for the NMT model in recent years, its performance on low-resource language pairs remains suboptimal compared to its high-resource counterparts due to the unavailability of large parallel corpora. Therefore, the implementation of NMT methods for LRLs pairs has been receiving the spotlight recently, leading to substantial research on this topic.Despite the availability of the pre-trained language models (PLMs), for small and medium-sized industries with insufficient hardware, there are many limitations in servicing the latest PLMs based NLP application software due to slow speed and insufficient memory, which may make it impossible. Because these techniques typically necessitate large amounts of data, they are much more difficult to service with PFA, particularly for low-resource languages. For this reason, the development of new methods for processing low-resource language pairs is still in its infancy, but researchers are beginning to explore alternative approaches such as transfer learning, multi-task learning, and pre-training that can help improve the NMT for downstream tasks with a small amount of data. By exploring the idea of PLMs, can we train on different LRLs (e.g., South African Language) to perform well on downstream tasks such as classification.

To solve this problem, this paper uses PLMs like AfriBerta (Ogueji et al., 2021), Afro-XLMR (Alabi et al., 2022) and AfroLM (Dossou et al., 2022) to run an experiment to improve the performance of NLP applications without changing the model through data pre- and post-processing, which is usually done in machine translation. The goal is to investigate how PLMs can be used to see if they could help LRL pairs do better in NMT.

2 METHODOLOGY

In this paper, three multilingual PLMs (AfriBERTa , AfroLM and Afro-XLMR) were fine-tuned and evaluated on the downstream task of topic classification. These models were fine-tuned on 2 labelled datasets , an isiZulu News dataset (Madodonga et al., 2022; 2023), which contained 14 different classes and the ANTC- African News Topic Classification dataset (Alabi et al., 2022) which consisted of five African languages, namely, isiZulu, Lingala, Malagasy, Pidgin, and Somali.

The overall experimentation that was followed in this paper was divided into 3 steps : (i) Train the PLMs and get the baseline performance of each language , (ii) Investigate different ways to improve the models performance through modifying the data and the model hyperparameters and (iii) Evaluating the model's performances. The PLMs are evaluated and tested on the Natural language processing (NLP) task of news topic classification. For each of the datasets, the performance was evaluated according to the languages.

2.1 Baseline Performance

The three PLM's were trained on the two datasets using the default hyperparameters. The baseline performances were measured using the F1 score of each model according to the respective languages. The results are presented in Table 1. These were the final , and best results that were retrieved from each model through the determination of the optimal values for the models' hyperparameters.

Table 1: Baseline performance (F1 Score)

Dataset	**Language**	**AfriBERTa**	**Afro-XLMR**	**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

3 Experimentation and Evaluation

From the baseline results that were retrieved, experimentation was conducted in an attempt to improve the PLMs performance. This was carried out through 2 different methods : (i) Modification of the architecture and (ii) Modification of the data. All the experimentation was carried out only using the isiZulu News dataset. The results that were obtained are presented in Table 2.

Modification of the architecture was explored , through freezing the models' encoder layers (lower layers), except for the classifier layer (HuggingFace). The classifier layer was initialised and adjusted according to the labels of the isiZulu data. The results that were obtained are presented in Table 2. This modification however did not improve model performance and resulted in a decrease of the F1 score for all 3 PLMs.

Table 2: Experimentation performance (F1 Score)

Implemented Method	**AfriBERTa**	**Afro-XLMR**	**AfroLM**
Freezing lower layers	0.424	0.464	0.384
Using the data in top 10 classes	0.514	0.771	0.643
Augmented Data	0.538	0.724	0.591

4 Conclusion

Based on the results presented on the baseline performance, the models performed best in the classification tasks on languages they were originally trained on. This highlights the importance of the inclusion of "low-resourced" languages in training PLMs and further emphasises the value that could be derived from the application of these PLMs in down-stream tasks such as classification. With the inclusion of data modification, the performances of all the PLMs increased, however slightly which suggests the importance of data pre and post processing. To ensure improved results from this experiment, advanced data modification methods can be investigated. This include text augmentation which is not purely based on the random rearrangement of words, but rather lexical and context-based techniques.

5 Acknowledgements

We would like to acknowledge funding from the ABSA Chair of Data Science, the Google Research scholar program, TensorFlow Award for Machine Learning Grant and the NVIDIA Corporation hardware Grant. Many thanks to all the anonymous AfricaNLP reviewers for their valuable feedback.

References

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 4336–4349, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.382.

Bonaventure F. P. Dossou, Atnafu Lambebo Tonja, Oreen Yousuf, Salomey Osei, Abigail Oppong, Iyanuoluwa Shode, Oluwabusayo Olufunke Awoyomi, and Chris Chinenye Emezue. Afrolm: A self-active learning-based multilingual pretrained language model for 23 african languages, 2022. URL https://arxiv.org/abs/2211.03263.

HuggingFace. Training and fine-tuning. Online. URL https://huggingface.co/transformers/v4.2.2/training.html. Last accessed 03 February 2023.

Daniel Li, I Te, Naveen Arivazhagan, Colin Cherry, and Dirk Padfield. Sentence boundary augmentation for neural machine translation robustness. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7553–7557. IEEE, 2021.

Andani Madodonga, Vukosi Marivate, and Matthew Adendorff. IsiZulu News (articles and headlines) and Siswati News (headlines) Corpora - za-isizulu-siswati-news-2022, 10 2022. URL https://github.com/dsfsi/za-isizulu-siswati-news-2022.

Andani Madodonga, Vukosi Marivate, and Matthew Adendorff. Izindaba-tindzaba: Machine learning news categorisation for long and short text for isizulu and siswati. *Journal of the Digital Humanities Association of Southern Africa*, 4(01), Jan. 2023. doi: 10.55492/dhasa.v4i01.4449. URL https://upjournals.up.ac.za/index.php/dhasa/article/view/4449.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2006.07264*, 2020.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 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*, pp. 116–126, Punta Cana, Dominican Republic, nov 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.mrl-1.11. URL https://aclanthology.org/2021.mrl-1.11.