WE STILL DO **FINE** WITH **LESS** LABELS WHEN DOING NAMED ENTITY RECOGNITION ON AFRICAN LANGUAGES

Arnol Fokam
University of the Witwatersrand, Johannesburg, South Africa

**INTRODUCTION**

- Transformer models perform well on tasks such as Named Entity Recognition (NER) with African languages.
- While this is encouraging, in a low-resource setting, it would be advantageous to analyse the performance of models when the quality of the dataset used varies.

**OBJECTIVE**

- NER datasets consist of pairs of sentences. The first sentence in each pair is a sentence in any language.
- The other sentence consists of NER tokens that are labels for each word in the first sentence.
- In a low-resource setting, it is hard to find annotators that can provide labels for words in African Languages.
- Therefore, *How is the performance of our NER models affected by the availability of these labels for every sentences?*

**METHODOLOGY**

- We use 3 languages from the MasakhaNER [1] dataset.
  - Swahili
  - Nigerian-Pidgin
  - Kinyarwanda
- For each language, we construct derived dataset where the number of token labels per sentences is capped and the surplus removed.
- For each dataset created, we train a set of NER models and record the F1-score on an evaluation set left un-changed.
  - BERT
  - RoBERTa
  - Multilingual BERT

**RESULTS & FINDINGS**

- More labels per sentence does not necessarily mean more performance.
- NER models can surprisingly perform well with less labels.
- Multi-language models perform better in such scenarios.

**ANALYSIS**

- As we increase the cap from 1 to 10, the performance benefits reduces.
- There is still some margin of improvement on Nigerian Pidgin. Maybe due to its similarity with English which is one of the high-resource languages used during the pre-training of these NER models.

**CONCLUSION**

A linear increase in the number of labels per sentence does not forcefully lead to a consistent linear improvement in the performance of NER models on African Languages.

**REFERENCES**


**Acknowledgement:** This is a product of an NLP course taught by Jade Abbott, Co-Founder of Masakhane.