


Insights into Low-Resource Language Modelling: Improving Model Performances for South African Languages


Ruan Visser

(Stellenbosch University, Stellenbosch, South Africa)

 <https://orcid.org/0009-0003-3192-5890>, 21051410@sun.ac.za


Trieko Grobler

(Stellenbosch University, Stellenbosch, South Africa)

 <https://orcid.org/0000-0001-5274-0105>, tlgrobler@sun.ac.za

Marcel Dunaiski

(Stellenbosch University, Stellenbosch, South Africa)

 <https://orcid.org/0000-0003-1957-3979>, marceldunaiski@sun.ac.za

Abstract: To address the gap in natural language processing for Southern African languages, our paper presents an in-depth analysis of language model development under resource-constrained conditions. We investigate the interplay between model size, pretraining objectives, and multilingual dataset composition in the context of low-resource languages such as Zulu and Xhosa. In our approach, we initially pretrain language models from scratch on specific low-resource languages using a variety of model configurations, and incrementally add related languages to explore the effect of additional languages on the performance of these models. We demonstrate that smaller data volumes can be effectively leveraged, and that the choice of pretraining objective and multilingual dataset composition significantly influences model performance. Our monolingual and multilingual models, exhibit competitive, and in some cases superior, performance compared to established multilingual models such as XLM-R-base and AfroXLM-R-base.

Keywords: Language Modelling, Low-Resource Languages, Transformers, Multilingual, Pre-training

Categories: I.2.7, I.7.2

DOI: 10.3897/jucs.118889

1 Introduction

In recent years considerable advancements have been made in natural language processing (NLP) particularly in the field of language modelling. Language models are typically learnt through unsupervised tasks called pretraining on large unlabelled datasets, which is followed by fine-tuning where a model is adapted and refined to a specific downstream language task.

A substantial challenge arises when dealing with languages that lack the conventional data volume needed for pretraining. Conventional high-resource languages are typically trained with gigabytes, or in some cases terabytes, of text data [Zhuang et al. 2021, Martin et al. 2020, Xu et al. 2020, Souza et al. 2020]. To overcome this barrier for low-resourced languages a common approach is to fine-tune on existing pretrained

massively multilingual models (MMMs) [Kalyan et al. 2021, Tela and Woubie 2020]. However, MMMs are typically significantly larger than monolingual models, due to the larger vocabularies required to accommodate for multiple languages that often comprise different scripts. This larger size can make these models more difficult to deploy particularly in settings that are restricted in terms of available computational resources [Abdaoui et al. 2020, Zhao et al. 2021, Rocholl et al. 2021].

Recent studies have shown that the performance of specific languages using MMMs can be improved by continuing the pretraining task with additional data from the corresponding languages [Muller et al. 2021, Alabi et al. 2022, Chau and Smith 2021]. However, the problem of having large bulky vocabularies remains with continually pretrained MMMs [Kuratov and Arkhipov 2019, Adelani et al. 2022b, Arkhipov et al. 2019].

In contrast, other studies have shown that well-performing pretrained models may be obtained with significantly smaller amounts of text [Martin et al. 2020, Micallef et al. 2022, Micheli et al. 2020, Martin et al. 2022]. However, the performance of these models either trail behind continually pretrained models or are only compared to older MMM alternatives [Martin et al. 2022, Micallef et al. 2022, Gessler and Zeldes 2022].

In the context of pretraining configurations, two of the main elements are the pretraining objective which is the mechanism with which the model learns from unsupervised data, and model size which controls the number of parameters within a model. Most studies that focus on low-resource language modelling only evaluate the efficacy of a single pretraining objective, usually masked language modelling (MLM) [Devlin et al. 2019] and a single model size, usually the base-sized models [Martin et al. 2022, Micallef et al. 2022, Ralethe 2020, Ogueji et al. 2021, Haq et al. 2023, Parida et al. 2021]. Relatively little focus has been placed on the evaluation of already established efficient model pretraining objectives, such as the replace token detection (RTD) [Clark et al. 2020] objective, and the impact that different model sizes have on model performance for low-resource languages.

In this paper, we explore the efficacy of pretraining language models from scratch for a number of Southern African languages by focusing on model size selection, alternative pretraining objectives, and the impact of multilinguality on the performance of these models.

Our results show that smaller data volumes can be effectively utilized for pretraining and that better model configurations can improve model performances significantly when additional data is not available. For example, we find that a small model pretrained on Xhosa obtained better downstream performance than larger base-sized models, while requiring significantly less compute to train. Additionally, we highlight that the selection of the pretraining objective and the inclusion of multiple related languages can significantly influence the downstream performance of these models. More specifically, we find that the downstream performance of low-resource languages, such as Xhosa and Zulu, can be significantly improved with the addition of related languages when pretraining, given a sufficiently large model size and vocabulary.

2 Related Work

Language models have conventionally relied on vast amounts of raw textual data, as illustrated by models such as RoBERTa [Zhuang et al. 2021], which was pretrained on 160GB of English text. However, most languages have significantly less raw textual data that are publicly available compared to high-resource languages such as English. In

light of this limitation, the most common approach is to use MMMs that can be adapted for language understanding tasks in languages that do not have dedicated monolingual models.

Prominent MMMs such as XLM-R [Conneau et al. 2020] and mBERT [Devlin et al. 2019] are trained on 100 and 104 languages, respectively. They exhibit strong performances across a broad range of languages, even for languages unseen during the pretraining process [Conneau et al. 2018]. Despite the advances that these models offer for cross-lingual understanding problems, their performances fall short of most dedicated monolingual counterparts for two main reasons. Firstly, the inherent “curse of multilinguality”, which [Alabi et al. 2022] define as the trade-off between language coverage and model capacity, leads to degraded performances across all languages if a model is trained for more than a certain number of languages [Conneau et al. 2020]. Secondly, low-resource languages often suffer from inadequate representation in the vocabulary of multilingual models. This has the undesired effect of splitting words into multiple smaller subwords unnecessarily, which increases the input sequence length and therefore hampers the model from learning effectively [Kalyan et al. 2021].

Language adaptive fine-tuning (LAFT) is an approach to mitigate these challenges which has recently found more widespread adoption. LAFT starts with a MMM, such as XLM-R, which is then further pretrained on a single target language or more than one target languages, in the case of multilingual adaptive fine-tuning (MAFT). Monolingual models such as RuBERT [Kuratov and Arkhipov 2019], BERTimbau [Souza et al. 2020], and SlavicBERT [Arkhipov et al. 2019], as well as multilingual models such as AfroXLM-R [Alabi et al. 2022] show the efficacy of this approach for a wide range of different languages.

In contrast, some studies have challenged this practice and instead argue that language models may be trained from scratch even when significantly less text is available while obtaining comparable downstream performance. [Martin et al. 2020] demonstrated that pretraining BERT on only 4 GB of French CommonCrawl data can yield similar downstream performance compared to models trained on over 130 GB of text. Similarly, [Micheli et al. 2020] trained small BERT models on varying amounts of CommonCrawl French text and found that 100 MB of text was sufficient to pretrain a well-performing French question-answering model. Similarly, [Micallef et al. 2022] explored the impact of using different Maltese pretraining data volumes from various sources on multiple downstream applications. They found that their model which was pretrained on only 46 million tokens (250 MB) achieved comparable results to an mBERT model that was further pretrained and adapted for Maltese. Other studies have suggested to pretrain language models from scratch on languages from specific regions. For example, [Ogueji et al. 2021] show the efficacy of pretraining on a set of African languages from scratch with AfriBERTa. They illustrate how multilingual models can be pretrained on less than 1 GB of text while obtaining better downstream performance on a variety of tasks when compared to massively multilingual alternatives. Additionally, [Kakwani et al. 2020] created IndicBERT, a multilingual model pretrained from scratch on 12 Indian languages, and found that their models are competitive or even outperform their trained word embeddings on various benchmarks for Indian languages. Despite these promising attempts, it is still unclear whether language models pretrained from scratch with limited amounts of resources can obtain better downstream performance compared to studies that utilize LAFT or MAFT using the same data constraints.

2.1 Pretraining Objective

Masked language modelling (MLM) introduced by [Devlin et al. 2019] was used to pretrain BERT which obtained state-of-the-art performance on the General Language Understanding Evaluation (GLUE) benchmark [Wang et al. 2018]. For the MLM objective, roughly 15% of the tokens of the pretraining corpus are typically masked out and the model tasked with predicting these masked tokens in a sequence. Most studies on low-resource language modelling have focused on models using the MLM pretraining objective [Micheli et al. 2020, Micallef et al. 2022, Ogueji et al. 2021, Alabi et al. 2022, Martin et al. 2022, Samuel et al. 2023, Gessler and Zeldes 2022].

In terms of pretraining efficiency, a significant improvement was achieved by the introduction of the replace token detection (RTD) model objective [Clark et al. 2020]. ELECTRA was the first model to use the RTD objective instead of MLM [Clark et al. 2020]. In RTD, tokens in a sequence are replaced with similar generated tokens and the model is tasked to identify the replaced tokens. Unlike MLM which usually only predicts 15% of tokens, RTD predicts whether each token in the training data is replaced or not. [Clark et al. 2020] find that the RTD strategy used to pretrain ELECTRA enables it to substantially outperform BERT in the GLUE benchmark, given the same model size, data volume, and compute budget. They argue that the higher number of decisions made per sequence is the primary reason for this performance increase. ConvBERT [Jiang et al. 2020] further improves the efficiency of ELECTRA by using a span-based dynamic convolution operator within the self-attention mechanism, while using the same RTD pretraining objective. They find that, when pretrained with the English OpenWebText corpus [Gokaslan and Cohen 2019], ConvBERT-base outperforms ELECTRA-base on the GLUE benchmark while only using 25% of the pretraining cost. [Daðason and Loftsson 2022] also find that ConvBERT outperforms ELECTRA when pretrained on Icelandic data and evaluated on four downstream tasks, namely, named-entity recognition, part-of-speech tagging, dependency parsing, and automatic text summarization.

2.2 Model Size

Neural networks with a large number of parameters are more likely to overfit on smaller datasets [Srivastava et al. 2014, Defernez and Kemsley 1999, Horenko 2020, Tsigler and Bartlett 2022]. However, many studies on low-resource language modelling use relatively large base-sized models [Micallef et al. 2022, Ogueji et al. 2021, Martin et al. 2022, Gessler and Zeldes 2022] or slight variations in architectural configurations. For instance, [Ogueji et al. 2021] experiment with relatively small parameter values in a range from 60.1 million to 102.6 million to evaluate the effect that the number of layers and attention heads have on downstream performance. They observed that deeper models outperform shallower ones when pretrained on multiple African languages, however, these improvements diminished as model depth increased. While studies exist that evaluate significantly smaller models for limited pretraining data volumes [Micheli et al. 2020], only a small number analyse the impact of model sizes [Gessler and Zeldes 2022]. In addition to larger models being more likely to overfit on a small pretraining dataset, larger pretrained transformers tend to produce unstable results when fine-tuned on tasks with smaller training sets [Phang et al. 2018].

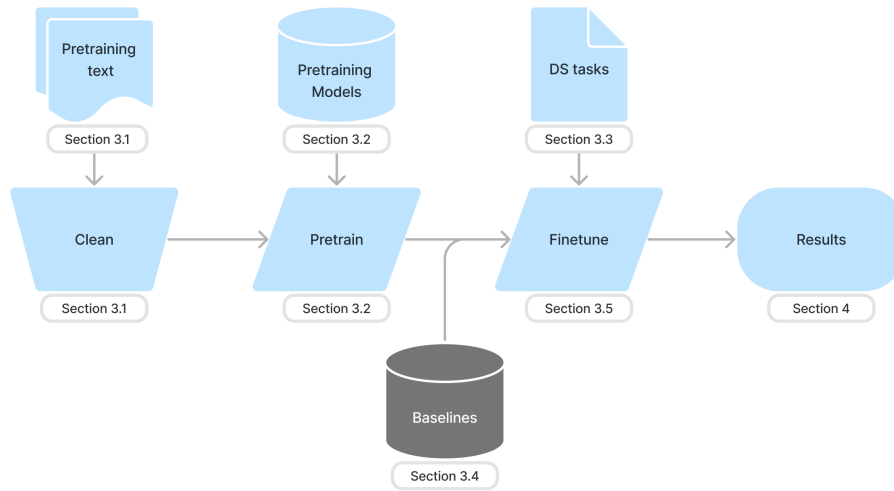


Figure 1: An illustration of the methodology we used for training and evaluating language models.

3 Methodology

Figure 1 illustrates the overall methodology used in this study. The following subsections provide a detailed discussion of each step. Section 3.1 covers our data collection and cleaning methodologies. Section 3.2 discusses the different models used and the pretraining process. Section 3.3 describes the various downstream tasks. Section 3.4 outlines the baseline models used. Lastly, Section 3.5 explains the fine-tuning process.

3.1 Pretraining Data

We pretrained our models on the language-specific subsets of the mC4 dataset [Xue et al. 2020], corresponding to our target languages: Xhosa, Zulu, Swahili, Nyanja, and Shona. Additionally, we included the scraped web content from Isolezwe¹ when pretraining models for Xhosa and Zulu, Voice of America (VOA) for Shona and Swahili, and lastly for Nyanja we used news articles from the Artificial Intelligence for Development initiative [Adelani et al. 2022a, Siminyu et al. 2021, Palen-Michel et al. 2022]. We keep approximately 5 MB of the textual data as a validation set for each language. Table 1 shows the amounts of textual data used to pretrain each of our models along with the language-specific amounts used to pretrain well-known multilingual models.

We found that many sentences or paragraphs within mC4 are boilerplate texts that do not contain any of the target language. To reduce the noise in the Zulu texts, we identified the 50 most occurring words in the Zulu Isolezwe corpus and remove sequences in Zulu mC4 that do not contain any of these words. This resulted in a significant reduction in pretraining text volume from 718 MB to 291 MB. The cleaning process involved similar

¹ <https://www.isolezwe.co.za/>

| Model or languages trained on | Abbr. | Total text volume | Total number of words | Target languages' text volume | Vocab size |
|--|--------|-------------------|-----------------------|-------------------------------|------------|
| XLM-R-base ^a | - | > 1 TB | 108 B | 1700 MB | 250K |
| AfroXLM-R-base ^b | - | > 1 TB | 109 B | 4555 MB | 250K |
| Xho ^c | X | 204 MB | 23 M | 204 MB | 30K |
| Zul ^c | Z | 324 MB | 39 M | 324 MB | 30K |
| Swa ^c | S | 3070 MB | 483 M | 204 MB | 30K |
| Zul+Xho ^c | ZX | 528 MB | 62 M | 528 MB | 40K |
| Zul+Xho+Swa ^c | ZXS | 3598 MB | 545 M | 3598 MB | 40K |
| Zul+Xho+Swa _{300 MB} ^c | ZXS300 | 898 MB | 136 M | 898 MB | 40K |
| Zul+Xho+Swa+Sho+Nya ^c | All | ~4500 MB | 708 M | ~4500 MB | 50K |

^a Massively multilingual model (MMM)

^b Multilingual adaptive fine-tuning on MMM

^c Pretrained from Scratch

Table 1: Overview of the pretraining data volumes for baselines, XLM-R and AfroXLM-R, along with the data volumes we use to pretrain our models. We report the total data volumes, total number of words, and the volume of data belonging to our target languages.

steps for the other languages, where we identified and removed non-relevant content. In Table 2, we show the initial and cleaned data sizes for each language. Although there are more complex text cleaning methods, we found that this heuristic is a simple and effective method of removing irrelevant noisy text for low-resource languages.

3.2 Pretraining Configurations

We pretrained the different volumes described at the bottom of Table 2 using ELECTRA, ConvBERT and BERT. Additionally, we trained both small and base-sized models for each of the pretraining datasets and pretraining objectives. We used the Huggingface implementation of Wordpiece [Wu et al. 2016] for tokenization and set the vocabulary size to 30 522 for our monolingual models. For our multilingual models trained with larger numbers of pretraining languages we used larger vocabulary sizes. More specifically, for models pretrained on 2 or 3 languages we increase the vocabulary size to 40 522, and for models trained on 5 languages we increase it further to 50 522. We experimented with a larger 100 522 vocabulary size for our model that contain all 5 target languages, however, this did not result in significant downstream improvements. Lastly, to ensure that lower-resource languages are not underrepresented within each of the multilingual vocabularies, we sampled an equal amount of pretraining text from each language when creating a vocabulary. Similar to IndicBERT, we limit the maximum sequence length to 128 tokens to reduce the computational requirements.

| Language | Original | Cleaned |
|----------|-------------|-------------|
| | Data Volume | Data Volume |
| Zulu | 718 MB | 291 MB |
| Xhosa | 175 MB | 125 MB |
| Swahili | 3036 MB | 2924 MB |
| Shona | 541 MB | 319 MB |
| Nyanja | 387 MB | 279 MB |

Table 2: Comparison of CommonCrawl (mC4) dataset sizes before and after cleaning.

Our pretrained models can be accessed at <https://doi.org/10.5281/zenodo.12686740>.

3.3 Downstream tasks

In addition to limited pretraining data, low-resource languages often also do not have language-specific downstream task. Recently, a handful of fine-tuning datasets for African languages were published by the Masakhane project². Three of these datasets are MansakaNER2.0 [Adelani et al. 2022b] for named entity recognition (NER), MansakaPOS [Dione et al. 2023] for part of speech tagging (POS), and MasakhaNEWS [Adelani et al. 2023] for news topic classification.

MansakaNER2.0 contains sentences from scraped local news sources for each of our target languages where each instance was labelled with named-entities, identified by three native speakers of the respective target language. In total, there are between 4 800 and 11 000 annotated sequences for each language in the dataset. MasakhaPOS was constructed from the same scraped data as MansakaNER2.0. However, with approximately 1500 sentences used per language, this dataset is significantly smaller than MasakaNER2.0. The annotation process also differs compared to MasakaNER2.0. In order to avoid the tedious and expensive annotation process, only 100 sentences were manually annotated with associated parts-of-speech [Dione et al. 2023]. Following this, a RemBERT model was trained on these sentences and then used to predict annotations for the remaining sentences. Subsequently, annotators were tasked with correcting erroneous predictions made by RemBERT.

For MasakhaNEWS, the annotation process is similar to MasakhaPOS. In the first stage annotators manually labelled the topics of 200 articles. In the second stage, these manually annotated articles were combined with others having predefined labels, obtained from the source websites, and used to fine-tune AfroXLM-R-base. This fine-tuned model was then used to predict the remaining articles. Again, annotators were asked to correct any mistakes made by the classifier. The size of MasakhaNEWS is comparable to that of MasakhaPOS, containing a similar number of observations. We list the training, validation and test distributions of each downstream tasks in Table 3.

² <https://www.masakhane.io/>

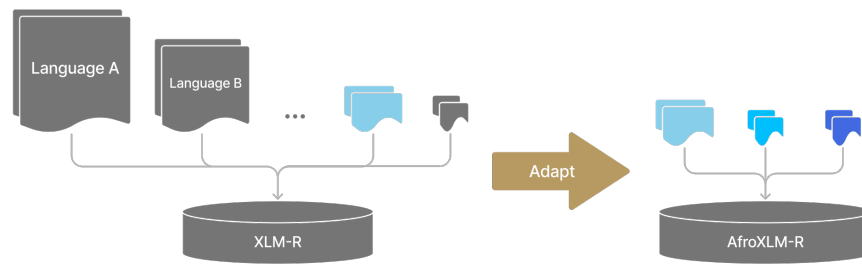


Figure 2: Illustration of multilingual pretraining from scratch on the left and language adaptive pretraining on the right. Multilingual models pretrained from scratch, such as XLM-R and our multilingual models, start with randomly initialized model weights which are then pretrained on multiple languages simultaneously. In language adaptive pretraining a multilingual model, such as XLM-R, is continuously pretrained on one or more target languages. For example, to create AfroXLM-R, XLM-R was further pretrained using 17 African languages.

| Language | Data source | MasakhaNER2.0 data splits | MasakhaPOS data splits | MasakhaNEWS data splits |
|---------------|-------------|---------------------------|------------------------|-------------------------|
| Swahili (swa) | VOA Swahili | 6593/ 942/ 1883 | 675/ 134/ 539 | 1658/ 237/ 476 |
| Xhosa (xho) | Isolezwe | 5718/ 817/ 1633 | 752/ 150/ 601 | 1032/ 147/ 297 |
| Zulu (zul) | Isolezwe | 5848/ 836/ 1670 | 753/ 150/ 601 | N/A |

Table 3: Overview of data sources and the distribution in training, development, and test splits for each downstream dataset, categorized by language.

3.4 Baselines

XLM-RoBERTa-base [Conneau et al. 2020] was pretrained using CC-100, a Common-Crawl dataset comprising 100 different languages. This includes two of our target language (Xhosa and Swahili) and six other African languages. XLM-R-base is trained on significantly more pretraining text compared to mBERT [Devlin et al. 2019] and is based on RoBERTa [Zhuang et al. 2021]. RoBERTa improves on the original BERT by training with larger batch sizes, utilizing byte pair encoding [Shibata et al. 1999] instead of WordPiece, and exclusively pretraining with the MLM (Masked Language Model) objective. [Conneau et al. 2020] find that XLM-R-base outperforms mBERT on most cross-lingual benchmarks.

AfroXLM-R-base [Alabi et al. 2022] was created by further pretraining an XLM-R-base model on 17 African languages which includes all of our target and three higher-resource languages commonly spoken in Africa (English, French and Arabic). [Alabi et al. 2022] show that AfroXLM-R-base outperforms XLM-R-base on each of the language subsets

within MasakhaNER2.0. Figure 2 illustrates the pretraining process of both baseline models.

3.5 Fine-Tuning

We fine-tune each of the pretrained models using the same control parameters used by [Clark et al. 2020]. However, we do not use layer-wise learning rate decay and instead use smaller learning rates similar to [Adelani et al. 2022b]. Accordingly, we use a learning rate of $2e-4$ for our small models and $5e-5$ for our base models. We fine-tune each pretrained model for 20 epochs on each dataset. To verify the robustness of our results, each model was fine-tuned and evaluated 15 times on each downstream task.

4 Results

4.1 Pretraining Results

In this section we show the results of analyzing the pretraining accuracy of ELECTRA models across different pretraining datasets, focusing on aspects such as a model’s capacity, the number of languages used during pretraining, the relative proportions of each language in multilingual datasets, as well as the impact that the pretraining duration on language-specific model has on performance. We used Isolezwe news datasets for Zulu and Xhosa, and VOA for Swahili, with a portion of these datasets reserved for validation. To mitigate the impact of outliers and noise, we applied a rolling mean with a window size of three observations for smoothing.

Figure 4 shows the achieved accuracy of monolingual and multilingual ELECTRA-base models during the pretraining process, using both the MLM and the RTD objectives. This figure contrasts training, shown as the dashed lines, and validation accuracies as solid lines across different languages. We observe that monolingual models, especially for lower-resource languages such as Zulu and Xhosa, are more prone to overfitting compared to multilingual models. This can be seen by the decline of the RTD validation accuracy for the Xhosa monolingual model after 100,000 training steps. A similar but less severe overfitting issue can be observed for the Zulu model after 200,000 steps. Including more languages within the pretraining dataset enhances pretraining accuracy of Zulu and Xhosa, as shown by the narrowing gap between validation and training RTD accuracies, and an overall increase in validation accuracy. Interestingly, we observe the opposite trend for the models that use the MLM objective where the incorporation of additional languages leads to worse MLM accuracy. It is unclear what causes this behaviour. We hypothesise that the improvement in RTD validation accuracy in multilingual models could be due to the models’ worse performing generators, which in turn could make the RTD task easier.

Interestingly, when Swahili, a higher-resource language, dominates the pretraining dataset (grey and black lines in Figure 3 and Figure 4), the language specific performance is similar irrespective of the amount of languages in the pretraining dataset. This can be seen on the Swahili pretraining results, and the grey and black lines in Figure 3 and Figure 4.

When comparing the performance of smaller and base-sized models, it becomes evident that base-sized models significantly outperform their smaller counterparts, particularly in MLM tasks. Here, generators for small models reach a maximum accuracy of 0.5, whereas generators for base models achieve accuracies as high as 0.67. Similarly,

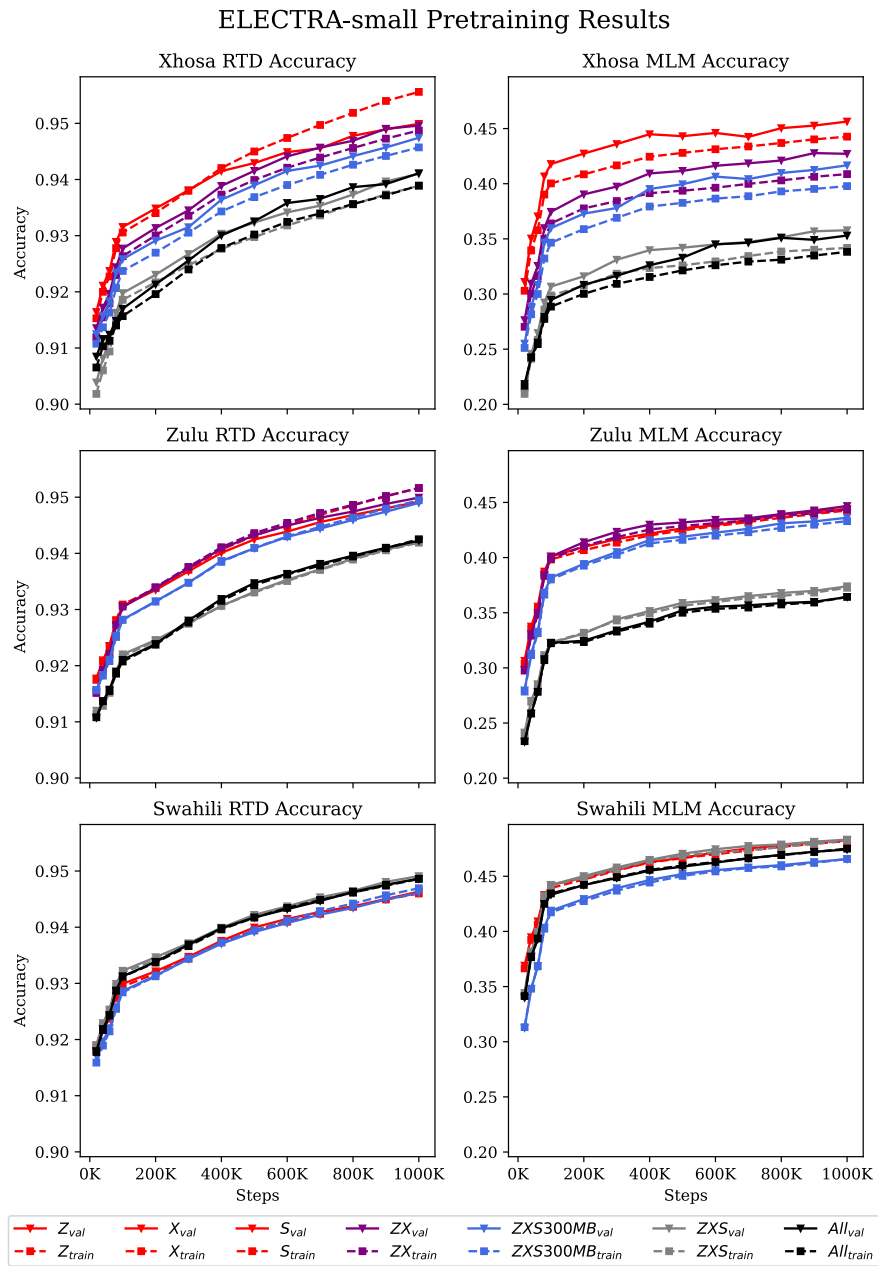


Figure 3: Language-specific pretraining MLM and RTD accuracy as we increase the number of pretraining steps for ELECTRA-small models on various monolingual and multilingual datasets. Dashed lines represent training accuracy, while solid lines indicate validation accuracy.

ELECTRA-base Pretraining Results

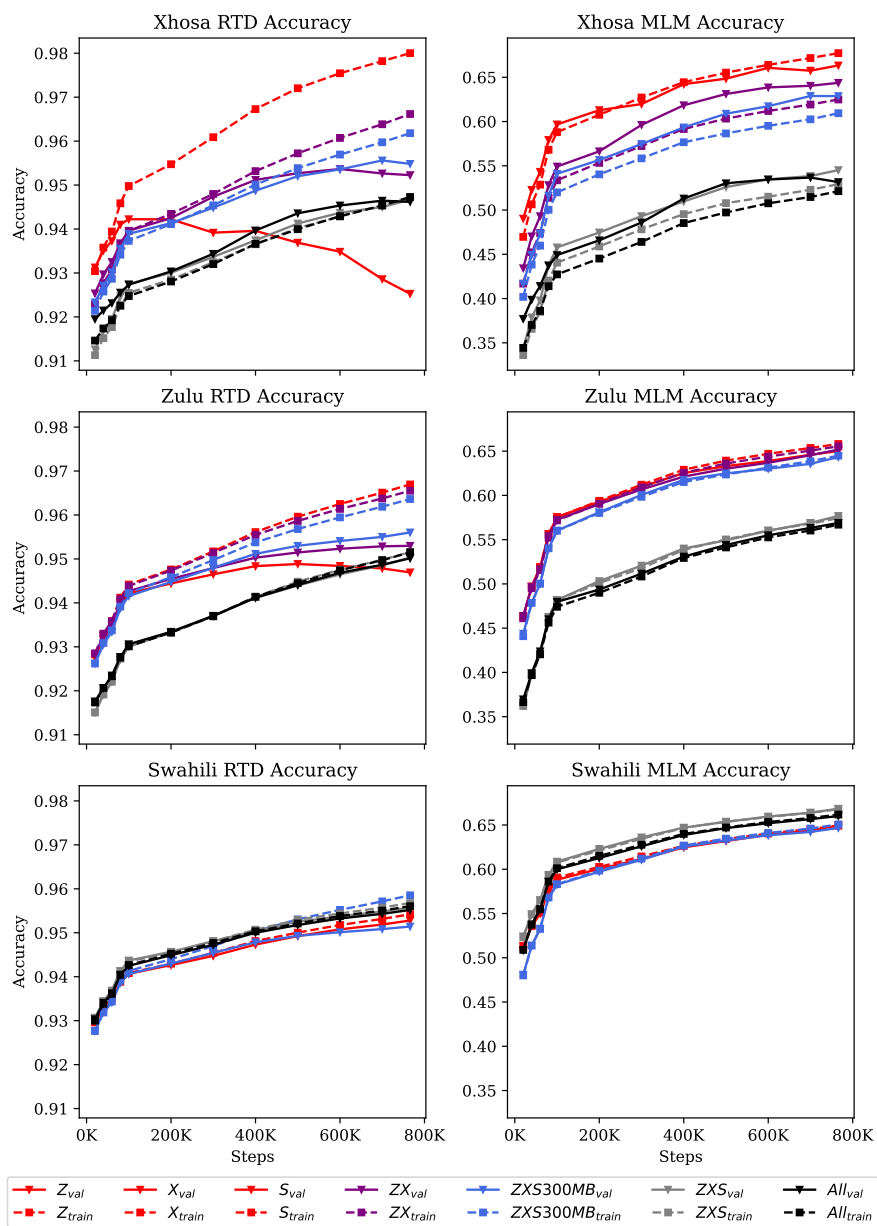


Figure 4: Language-specific pretraining MLM and RTD accuracies as the number of pretraining steps is increased for ELECTRA-base models on various monolingual and multilingual datasets.

RTD performance is also better in multilingual base-sized models compared to smaller ones. However, it should be noted that smaller models, despite their limitations in MLM tasks, are less prone to overfitting compared to base models.

4.2 Downstream Results

Figure 5 illustrates the impact of model size (small and base), model type (RTD and MLM), and the number of pretraining languages on the downstream task performance of named-entity recognition using the MasakhaNER2.0 dataset. We can see that for smaller models the use of additional pretraining languages generally leads to reduced performance. However, ConvBERT-small exhibits an exception, showing improved NER performance for Zulu and Xhosa when pretrained on both languages. In contrast, for base-sized models, NER performance generally improves with the addition of more pretraining languages. This could suggest base models can better adapt to linguistic diversity in pretraining datasets. In terms of model variants, we observe that the small-models that are pretrained using RTD, namely ELECTRA-small and ConvBERT-small, tend to perform better than BERT-small. However, for base-sized models, BERT and ConvBERT perform significantly better than ELECTRA models.

Figure 6 depicts the results when our models are evaluated using the downstream MasakhaPOS benchmark dataset, with the top row of graphs showing small-sized model performances and the bottom row the corresponding base-sized model performances. The results indicate that the Swahili POS performance of small models improves as the number of pretraining languages increases. However, there is a deterioration in performance for Zulu POS and the Xhosa POS tasks when Swahili is added to the pretraining dataset. We believe this could be attributed to the disproportionate size of Swahili dataset which makes up 85% of the ‘ZXS’ training data. Interestingly, the addition of Shona and Nyanja, each with approximately 300 MB of pretraining text, improves Zulu and Xhosa POS performances when our small models are used compared to pretraining only with Zulu, Xhosa and Swahili. Base-sized models generally improve the POS performances when more languages are used, but this trend reverses when expanding from three to five languages, indicating potential capacity dilution. In terms of model type, BERT-base tends to outperform the RTD models. However, in the case of the Xhosa POS task, base-sized models trained using RTD show more significant improvements when Zulu and Swahili are added to the pretraining dataset, leading to superior downstream performance compared to their multilingual base-sized MLM counterparts.

Figure 7 presents the fine-tuning results when the MasakhaNEWS downstream benchmark dataset is used for evaluation. In the case of small models, we observe a decline in performance as the number of pretraining languages increases. In contrast, for base-sized models, altering the number of languages in the pretraining dataset does not significantly affect news topic classification performances. Notably, among the base-sized models, BERT-base consistently outperforms our other models in news topic classification.

4.3 Baselines Comparisons

In Table 4 we compare the downstream performance of well-known multilingual models, XLM-R-base and AfroXLM-R-base (an XLM-R-base model adapted for African languages), along with our pretrained monolingual and multilingual models. Our model selection aligns with trends identified in Figure 5 through 7. For instance, we chose

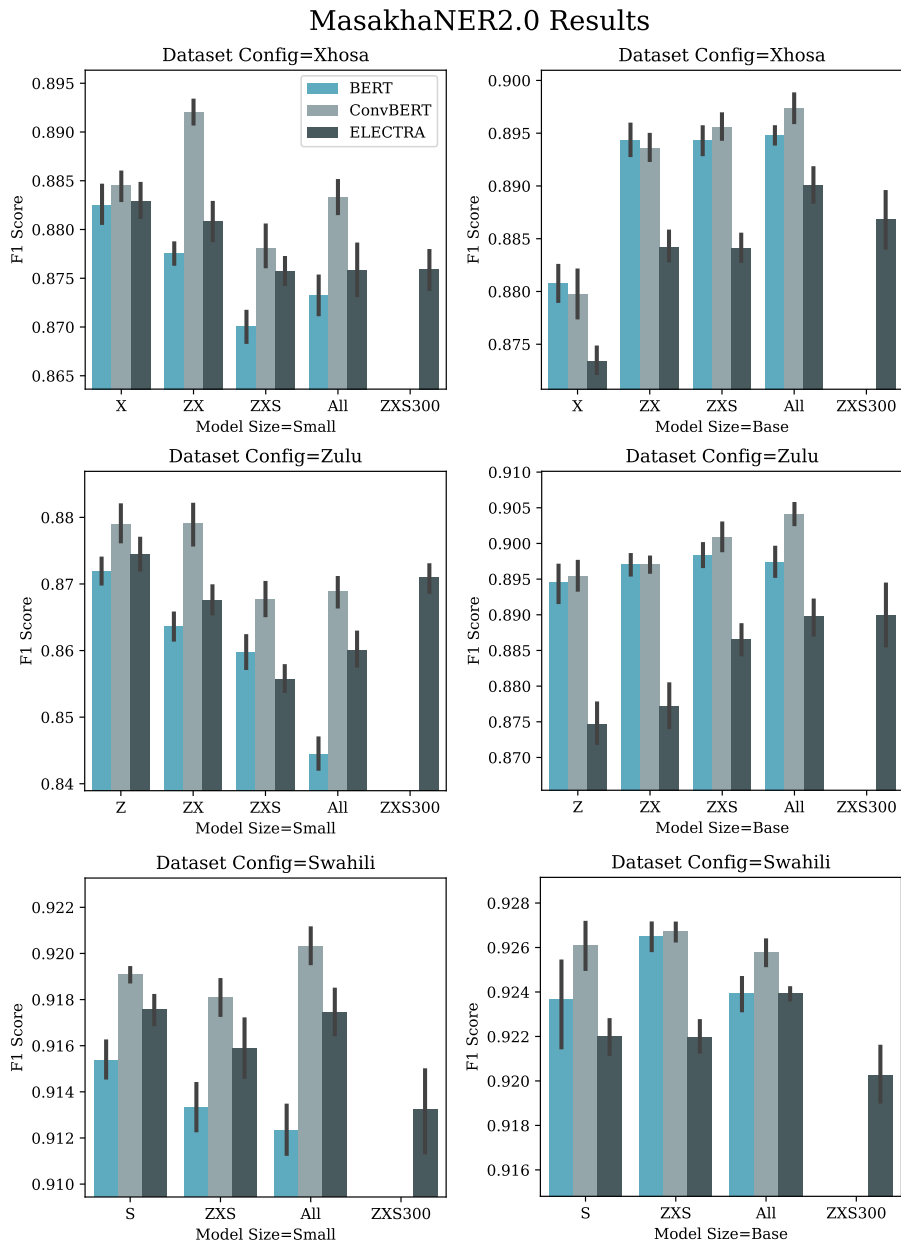


Figure 5: The impact of model size, model type and number of pretraining languages on MasakhaNER2.0 downstream F1 performance. The pretraining datasets evaluated are shown on the X-axes with each string corresponding to a the abbreviations in Table 2. The vertical line on top of each bar indicates the standard deviation.

MasakhaPOS Results

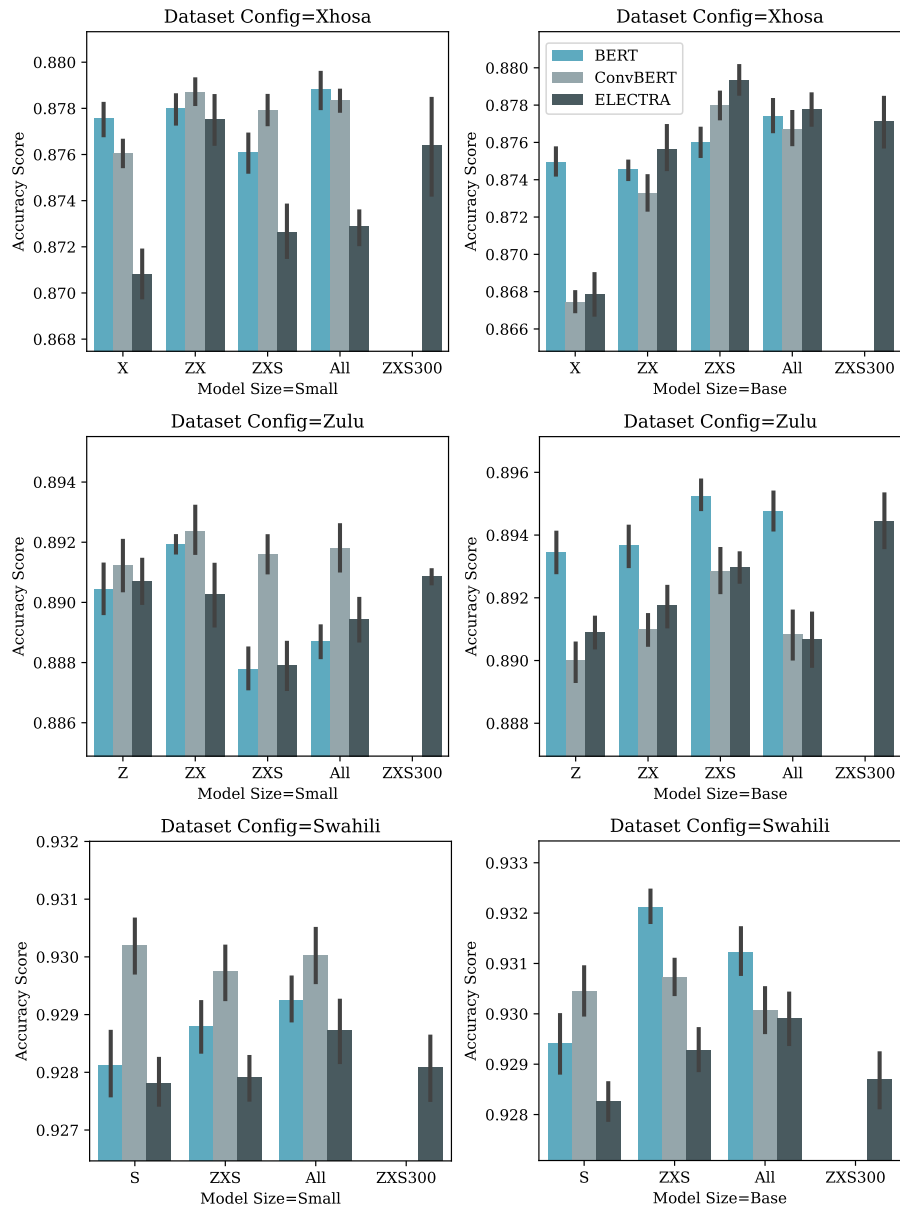


Figure 6: The impact of model size, model type and number of pretraining languages on MasakhaPOS downstream accuracy performance. The pretraining datasets evaluated are shown on the X-axes with each string corresponding to the abbreviations in Table 2. The vertical line on top of each bar indicates the standard deviation.

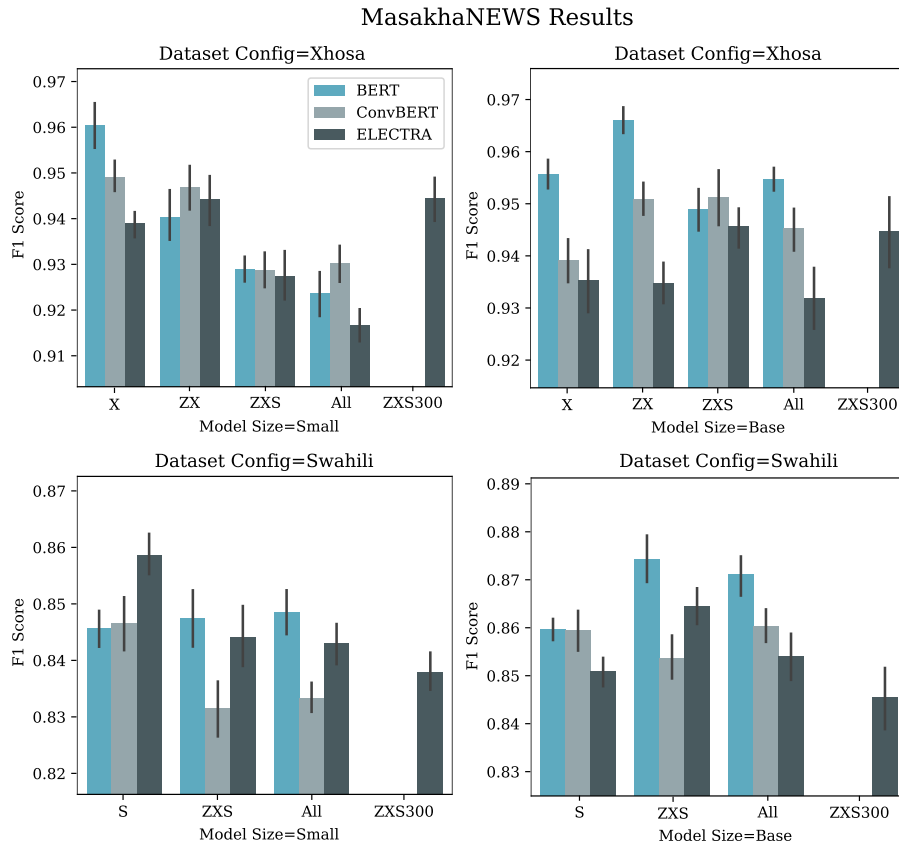


Figure 7: The impact of model size, model type and number of pretraining languages on MasakhaNEWS downstream F1 performance. The pretraining datasets evaluated are shown on the X-axis with each string corresponding to a the abbreviations in Table 2. The vertical line on top of each bar indicates the standard deviation.

ConvBERT-small for all our Xhosa monolingual results, as smaller models showed better performance in lower-resource pretraining scenarios. Our multilingual models use the extensive ‘All’ (containing Zulu, Xhosa, Swahili, Nyanja and Shona) dataset for NER, reflecting their improved performance with increased language diversity. Additionally, we use ConvBERT-base models when reporting monolingual Zulu, Swahili, and all multilingual NER results. However, for POS and news topic classification, BERT-base models were selected due to their superior performance.

[Ogueji et al. 2021] also pretrained BERT language models from scratch on multiple low-resource African languages. They evaluate their models across more than ten language using NER and find that their largest model, AfroBERTa-large, obtains similar performance to XLM-R-base with only a 0.06% different in average accuracy. Similarly, [Kakwani et al. 2020] compare their multilingual IndicBERT-base model against XLM-R-base on a wide variety of downstream tasks with each task covering approxi-

| | POS | NER | NEWS |
|----------------|----------------------|----------------------|----------------------|
| <i>Xhosa</i> | | | |
| XLM-R | 0.874 (0.002) | 0.873 (0.003) | 0.838 (0.049) |
| AfroXLM-R | 0.885 (0.002) | 0.858 (0.080) | 0.932 (0.013) |
| Monolingual | 0.876 (0.001) | 0.884 (0.003) | 0.951 (0.007) |
| Multilingual | 0.876 (0.002) | 0.897 (0.003) | 0.956 (0.005) |
| <i>Zulu</i> | | | |
| XLM-R | 0.881 (0.002) | 0.850 (0.004) | N/A |
| AfroXLM-R | 0.895 (0.001) | 0.885 (0.005) | N/A |
| Monolingual | 0.894 (0.001) | 0.895 (0.006) | N/A |
| Multilingual | 0.895 (0.003) | 0.904 (0.003) | N/A |
| <i>Swahili</i> | | | |
| XLM-R | 0.929 (0.001) | 0.922 (0.001) | 0.860 (0.009) |
| AfroXLM-R | 0.930 (0.001) | 0.924 (0.003) | 0.876 (0.008) |
| Monolingual | 0.929 (0.001) | 0.926 (0.002) | 0.861 (0.005) |
| Multilingual | 0.932 (0.001) | 0.926 (0.001) | 0.870 (0.009) |

Table 4: Downstream evaluation of our models compared to baselines on MasakhaNER2.0 (NER), MasakhaPOS (POS) and MasakhaNEWS (NEWS), for Xhosa, Zulu and Swahili. For NER and NEWS we report F1 macro performance and for POS we report accuracy. The reported performance of each model is averaged over 15 finetuning runs.

mately 10 or more Indian languages. For article genre classification their model does 0.31% better on average across all languages; however, for NER, IndicBERT-base does 0.54% worse than XLM-R-base. These small margins indicate that IndicBERT-base’s and AfroBERTa-large’s performances are on par with XLM-R-base on average.

Our monolingual models outperform XLM-R-base which suggests that effective low-resource language models can be pretrained with limited data. Additionally, our multilingual models often outperform AfroXLM-R-base in various tasks, with exceptions in Xhosa POS and Swahili news topic classification. Furthermore, our multilingual results consistently outperform XLM-R-base, and for certain tasks, such as Xhosa NEWS classification, by more than 11% accuracy. These outcomes highlight the potential of custom-tailored pretraining strategies, especially when benchmarked against established multilingual models in the domain of low-resource language processing.

5 Discussion

5.1 Overfitting and Multilingual Pretraining Dynamics

A key observation is the tendency of low-resource monolingual models to overfit, which is particularly evident in the difference between the training and validation pretraining

results for monolingual Xhosa and Zulu ELECTRA-base models. Similar to results reported by [Conneau et al. 2020], we observe that the addition of languages in the pretraining dataset can improve the downstream results for our lower-resource languages, such as Zulu and Xhosa, for our base-sized models. Interestingly, we find that combining both of these low-resource languages led to less overfitting while also improving the downstream performance on each language. Furthermore, we find that the addition of Swahili leads to similar or better Zulu and Xhosa downstream results, even though this addition degrades both the RTD and MLM pretraining performance. This could suggest that more pretraining text could be beneficial for downstream results in low-resource languages even if it reduces pretraining performance in the target languages.

5.2 The influence of Model Size

In our study, smaller models showed less overfitting in low-resource languages such as Zulu and Xhosa. This is likely due to the fact that these models comprised fewer parameters that may lead to better generalization from the limited data. However, after we introduced additional languages, the base-sized models began to improve significantly, which suggests that they are able to handle more complex data better. However, there exists a limit since we find that adding too many languages leads to ‘capacity dilution’ which reduces model performances. This is similar to the degradation in overall model performance others have observed when increasing the number of languages a model is trained on [Conneau et al. 2020, Tan et al. 2018]. Therefore it is important to balance model size with training data characteristics and linguistic diversity. More specifically, smaller models are preferable for pretraining on limited monolingual data while larger models should be chosen for more diverse and larger dataset.

5.3 Model type and downstream task

From our analysis on the impact of different model types on downstream results, we find that small-sized RTD models tend to perform better when tasked with NER compared to small MLM models. ConvBERT-base, which uses RTD objective during training, performs the best for this downstream task. However, for POS and news topic classification BERT-base, the MLM models seem to perform better. However, related studies that also compare the performance differences between RTD and MLM models on similar downstream tasks have found the opposite results. For example, [Daðason and Loftsson 2022] find that ELECTRA and ConvBERT perform better than BERT on both Icelandic POS and NER, and the creators of AraELECTRA [Antoun et al. 2021] find that ELECTRA-base performs better on average compared to BERT-base on both NER and sentiment classification.

However, both these examples are pretrained on significantly more pretraining data (more than 50 GB), in comparison with our largest pretraining monolingual dataset, Swahili, which contains only 3 GB of pretraining data. This could suggest that RTD performs better than MLM when pretrained using more text and might only be more computationally efficient when sufficient pretraining data is available.

5.4 Limitations and Future Directions

The pretraining dataset’s composition, particularly the dominance of certain languages, may impact the generalizability of our findings. Future research should explore a wider

array of languages and more balanced datasets. Additionally, other pretraining objectives should be investigated and a larger variety of downstream tasks will yield better insights into efficient strategies for low-resource language modelling. We observed that the pretraining data may be very noisy, which can result in both worse pretraining accuracy and downstream performances.

We find that the performances between models can vary across tasks, with some downstream tasks favouring language-adaptive finetuning while pretraining from scratch performs better for others. For future research one could consider using hybrid or ensemble models that incorporate both to improve accuracy for low-resource languages.

6 Conclusion

Our study offers insights into pretraining language models for Southern African low-resource languages by highlighting the balance between model size, pretraining objectives, and multilingual dataset composition. We find that smaller models are less prone to overfitting in monolingual settings, particularly for low-resource languages such as Zulu and Xhosa. However, as language diversity increases in pretraining datasets, larger base-sized models demonstrate superior adaptability, albeit with a threshold beyond which performance starts to decline due to capacity dilution.

Interestingly, the addition of higher-resource languages, such as Swahili, in multilingual pretraining did not necessarily degrade language-specific performance which suggests that a larger corpus size can offset the potential dilution of language-specific capacity. This finding challenges the conventional notion of the “curse of multilinguality” and underscores the importance of balanced and diverse pretraining data for low-resource languages.

Moreover, our models, both monolingual and multilingual, displayed competitive, and in several instances, superior performance compared to established multilingual models such as XLM-R-base and AfroXLM-R-base. This underscores that it is feasible to develop effective language models with relatively small data volumes and presents a viable pathway for low-resource language processing.

Acknowledgements

This work was supported by Google’s TPU Research Cloud program.

References

- [Abdaoui et al. 2020] Abdaoui, A., Pradel, C. & Sigel, G.: “Load What You Need: Smaller Versions of Multilingual BERT.” *Proceedings Of SustaiNLP: Workshop On Simple And Efficient Natural Language Processing*. pp. 119-123 (2020,11).
- [Adelani et al. 2022a] Adelani, D., Alabi, J., Fan, A., Kreutzer, J., Shen, X., Reid, M., Ruiter, D., Klakow, D., Nabende, P., Chang, E., Gwadabe, T., Sackey, F., Dossou, B., Emezue, C., Leong, C., Beukman, M., Muhammad, S., Jarso, G., Yousuf, O., Niyongabo Rubungo, A., Hacheme, G., Wairagala, E., Nasir, M., Ajibade, B., Ajayi, T., Gitau, Y., Abbott, J., Ahmed, M., Ochieng, M., Aremu, A., Ogayo, P., Mukiibi, J., Ouoba Kabore, F., Kalipe, G., Mbaye, D., Tapo, A., Memdjokam Koagne, V., Munkoh-Buabeng, E., Wagner, V., Abdulmumin, I., Awokoya, A., Buzaaba, H., Sibanda, B., Bukula, A. & Manthalu, S.: “A Few Thousand Translations Go a Long Way! Leveraging Pre-trained Models for African News Translation.” *Proceedings Of The 2022 Conference Of The North American Chapter Of The Association For Computational Linguistics: Human Language Technologies*. pp. 3053-3070 (2022,7).

- [Adelani et al. 2022b] Adelani, D., Neubig, G., Ruder, S., Rijhwani, S., Beukman, M., Palen-Michel, C., Lignos, C., Alabi, J., Muhammad, S., Nabende, P., Dione, C., Bukula, A., Mabuya, R., Dossou, B., Sibanda, B., Buzaaba, H., Mukiibi, J., Kalipe, G., Mbaye, D., Taylor, A., Kabore, F., Emezue, C., Aremu, A., Ogayo, P., Gitau, C., Munkoh-Buabeng, E., Memdjokam Koagne, V., Tapo, A., Macucwa, T., Marivate, V., Elvis, M., Gwadabe, T., Adewumi, T., Ahia, O., Nakatumba-Nabende, J., Mokono, N., Ezeani, I., Chukwuneke, C., Oluwaseun Adeyemi, M., Hacheme, G., Abdulmumin, I., Ogundepo, O., Yousuf, O., Moteu, T. & Klakow, D.: "MasakhaNER 2.0: Africa-centric Transfer Learning for Named Entity Recognition." *Proceedings Of The 2022 Conference On Empirical Methods In Natural Language Processing*. pp. 4488-4508 (2022,12).
- [Adelani et al. 2023] Adelani, D., Masiak, M., Azime, I., Alabi, J., Tonja, A., Mwase, C., Ogundepo, O., Dossou, B., Oladipo, A., Nixdorf, D., Emezue, C., Azzawi, S., Sibanda, B., David, D., Ndolela, L., Mukiibi, J., Ajayi, T., Ngoli, T., Odhiambo, B., Owodunni, A., Obiefuna, N., Muhammad, S., Abdullahi, S., Yigezu, M., Gwadabe, T., Abdulmumin, I., Bame, M., Awoyomi, O., Shode, I., Adelani, T., Kailani, H., Omotayo, A., Adeeko, A., Abeeb, A., Aremu, A., Samuel, O., Siro, C., Kimotho, W., Ogbu, O., Mbonu, C., Chukwuneke, C., Fanijo, S., Ojo, J., Awosan, O., Guge, T., Sari, S., Nyatsine, P., Sidume, F., Yousuf, O., Oduwale, M., Kimanuka, U., Tshinu, K., Diko, T., Nxakama, S., Johar, A., Gebre, S., Mohamed, M., Mohamed, S., Hassan, F., Mehamed, M., Ngabire, E. & Pontus Stenertorp: "MasakhaNEWS: News Topic Classification for African languages." *ArXiv*. (2023).
- [Alabi et al. 2022] Alabi, J., Adelani, D., Mosbach, M. & Klakow, D.: "Adapting Pre-trained Language Models to African Languages via Multilingual Adaptive Fine-Tuning." *Proceedings Of The 29th International Conference On Computational Linguistics*. pp. 4336-4349 (2022,10).
- [Antoun et al. 2021] Antoun, W., Baly, F. & Hajj, H.: "AraELECTRA: Pre-Training Text Discriminators for Arabic Language Understanding." *Proceedings Of The Sixth Arabic Natural Language Processing Workshop*. pp. 191-195 (2021,4).
- [Arkipov et al. 2019] Arkipov, M., Trofimova, M., Kuratov, Y. & Sorokin, A.: "Tuning Multilingual Transformers for Language-Specific Named Entity Recognition." *Proceedings Of The 7th Workshop On Balto-Slavic Natural Language Processing*. pp. 89-93 (2019,8).
- [Bender et al. 2021] Bender, E., Gebru, T., McMillan-Major, A. & Shmitchell, S.: "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?." *Proceedings Of The 2021 ACM Conference On Fairness, Accountability, And Transparency*. pp. 610-623 (2021).
- [Chau and Smith 2021] Chau, E. & Smith, N.: "Specializing Multilingual Language Models: An Empirical Study." *Proceedings Of The 1st Workshop On Multilingual Representation Learning*. pp. 51-61 (2021,11).
- [Chen and Manning 2014] Chen, D. & Manning, C.: "A Fast and Accurate Dependency Parser using Neural Networks." *Proceedings Of The 2014 Conference On Empirical Methods In Natural Language Processing (EMNLP)*. pp. 740-750 (2014,10).
- [Chiguvare and Cleghorn 2021] Chiguvare, P. & Cleghorn, C.: "Improving transformer model translation for low resource South African languages using BERT." *2021 IEEE Symposium Series On Computational Intelligence (SSCI)*. pp. 1-8 (2021).
- [Clark et al. 2020] Clark, K., Luong, M., Le, Q. & Manning, C.: "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators." *8th International Conference On Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. (2020).
- [Collobert et al. 2011] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K. & Kuksa, P.: "Natural Language Processing (Almost) from Scratch." *J. Mach. Learn. Res.* 12, 2493-2537 (2011,11).
- [Conneau et al. 2018] Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H. & Stoyanov, V.: "XNLI: Evaluating Cross-lingual Sentence Representations." *Proceedings Of The 2018 Conference On Empirical Methods In Natural Language Processing*. pp. 2475-2485 (2018).

- [Conneau et al. 2020] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L. & Stoyanov, V.: “Unsupervised Cross-lingual Representation Learning at Scale.” *Proceedings Of The 58th Annual Meeting Of The Association For Computational Linguistics*. pp. 8440-8451 (2020,7).
- [Daðason and Loftsson 2022] Daðason, J. & Loftsson, H.: “Pre-training and Evaluating Transformer-based Language Models for Icelandic.” *Proceedings Of The Thirteenth Language Resources And Evaluation Conference*. pp. 7386-7391 (2022,6).
- [Defernez and Kemsley 1999] Defernez, M. & Kemsley, E.: “Avoiding overfitting in the analysis of high-dimensional data with artificial neural networks (ANNs).” *The Analyst*. 124 11 pp. 1675-81 (1999).
- [Devlin et al. 2019] Devlin, J., Chang, M., Lee, K. & Toutanova, K.: “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” *Proceedings Of The 2019 Conference Of The North American Chapter Of The Association For Computational Linguistics: Human Language Technologies, Volume 1 (Long And Short Papers)*. pp. 4171-4186 (2019,6).
- [Dione et al. 2023] Dione, C., Adelani, D., Nabende, P., Alabi, J., Sindane, T., Buzaaba, H., Muhammad, S., Emezue, C., Ogayo, P., Aremu, A., Gitau, C., Mbaye, D., Mukiibi, J., Sibanda, B., Dossou, B., Bukula, A., Mabuya, R., Tapo, A., Munkoh-Buabeng, E., Memdjokam Koagne, V., Ouoba Kabore, F., Taylor, A., Kalipe, G., Macucwa, T., Marivate, V., Gwadabe, T., Elvis, M., Onyenwe, I., Atindogbe, G., Adelani, T., Akinade, I., Samuel, O., Nahimana, M., Musabeyezu, T., Niyomutabazi, E., Chimhenga, E., Gotosa, K., Mizha, P., Agbolo, A., Traore, S., Uchechukwu, C., Yusuf, A., Abdullahi, M. & Klakow, D.: “MasakhaPOS: Part-of-Speech Tagging for Typologically Diverse African languages.” *Proceedings Of The 61st Annual Meeting Of The Association For Computational Linguistics (Volume 1: Long Papers)*. pp. 10883-10900 (2023,7).
- [Dodge et al. 2020] Dodge, J., Ilharco, G., Schwartz, R., Farhadi, A., Hajishirzi, H. & Smith, N.: “Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping.” (2020).
- [Dodge et al. 2021] Dodge, J., Sap, M., Marasović, A., Agnew, W., Ilharco, G., Groeneveld, D., Mitchell, M. & Gardner, M.: “Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus.” *Proceedings Of The 2021 Conference On Empirical Methods In Natural Language Processing*. pp. 1286-1305 (2021,11).
- [Furman et al. 2021] Pérez, J., Furman, D., Alemany, L. & Luque, F.: “RoBERTuito: a pre-trained language model for social media text in Spanish.” *International Conference On Language Resources And Evaluation*. (2021).
- [Geiping and Goldstein 2022] Geiping, J. & Goldstein, T.: “Cramming: Training a Language Model on a Single GPU in One Day.” (2022).
- [Gessler and Zeldes 2022] Gessler, L. & Zeldes, A.: “MicroBERT: Effective Training of Low-resource Monolingual BERTs through Parameter Reduction and Multitask Learning.” *Proceedings Of The The 2nd Workshop On Multi-lingual Representation Learning (MRL)*. pp. 86-99 (2022,12).
- [Gokaslan and Cohen 2019] Gokaslan, A. & Cohen, V.: “OpenWebText Corpus.” (2019).
- [Haq et al. 2023] Haq, I., Qiu, W., Guo, J. & Tang, P.: “NLPashto: NLP Toolkit for Low-resource Pashto Language.” *International Journal Of Advanced Computer Science And Applications*. 14 (2023).
- [Hoffschmidt et al. 2020] D’Hoffschmidt, M., Belblidia, W., Heinrich, Q., Brendlé, T. & Vidal, M.: “FQuAD: French Question Answering Dataset.” *Findings Of The Association For Computational Linguistics: EMNLP 2020*. pp. 1193-1208 (2020,11).
- [Horenko 2020] Horenko, I.: “On a Scalable Entropic Breaching of the Overfitting Barrier for Small Data Problems in Machine Learning.” *Neural Computation*. 32, 1563-1579 (2020).
- [Howard and Ruder 2018] Howard, J. & Ruder, S.: “Universal Language Model Fine-tuning for Text Classification.” *Proceedings Of The 56th Annual Meeting Of The Association For Computational Linguistics (Volume 1: Long Papers)*. pp. 328-339 (2018,7).

- [Jiang et al. 2020] Jiang, Z., Yu, W., Zhou, D., Chen, Y., Feng, J. & Yan, S.: "ConvBERT: Improving BERT with Span-Based Dynamic Convolution." *Proceedings Of The 34th International Conference On Neural Information Processing Systems*. (2020).
- [Kakwani et al. 2020] Kakwani, D., Kunchukuttan, A., Golla, S., N.C., G., Bhattacharyya, A., Khapra, M. & Kumar, P.: "IndicNLP Suite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages." *Findings Of EMNLP*. (2020).
- [Kalyan et al. 2021] Kalyan, K., Rajasekharan, A. & Sangeetha, S.: "AMMUS : A Survey of Transformer-based Pretrained Models in Natural Language Processing." (2021).
- [Kim 2014] Kim, Y.: "Convolutional Neural Networks for Sentence Classification." *Proceedings Of The 2014 Conference On Empirical Methods In Natural Language Processing (EMNLP)*. pp. 1746-1751 (2014,10).
- [Kondratyuk and Straka 2019] Kondratyuk, D. & Straka, M.: "75 Languages, 1 Model: Parsing Universal Dependencies Universally." *Proceedings Of The 2019 Conference On Empirical Methods In Natural Language Processing And The 9th International Joint Conference On Natural Language Processing (EMNLP-IJCNLP)*. pp. 2779-2795 (2019,11).
- [Kreutzer et al. 2022] Kreutzer, J., Caswell, I., Wang, L., Wahab, A., Esch, D., Ulzii-Orshikh, N., Tapo, A., Subramani, N., Sokolov, A., Sikasote, C., Setyawan, M., Sarin, S., Samb, S., Sagot, B., Rivera, C., Rios, A., Papadimitriou, I., Osei, S., Suarez, P., Orife, I., Ogueji, K., Rubungo, A., Nguyen, T., Müller, M., Müller, A., Muhammad, S., Muhammad, N., Mnyakeni, A., Mirzakhlov, J., Matangira, T., Leong, C., Lawson, N., Kudugunta, S., Jernite, Y., Jenny, M., Firat, O., Dossou, B., Dlamini, S., Silva, N., Balli, S., Biderman, S., Battisti, A., Baruwu, A., Bapna, A., Baljekar, P., Azime, I., Awokoya, A., Ataman, D., Ahia, O., Ahia, O., Agrawal, S. & Adeyemi, M.: "Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets." *Transactions Of The Association For Computational Linguistics*. 10 pp. 50-72 (2022).
- [Kuratov and Arkhipov 2019] Kuratov, Y. & Arkhipov, M.: "Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language." (2019).
- [Liu et al. 2019] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V.: "RoBERTa: A Robustly Optimized BERT Pretraining Approach." (2019).
- [Martin et al. 2020] Martin, L., Muller, B., Ortiz Suárez, P., Dupont, Y., Romary, L., Clergerie, É., Seddah, D. & Sagot, B.: "CamemBERT: a Tasty French Language Model." *Proceedings Of The 58th Annual Meeting Of The Association For Computational Linguistics*. pp. 7203-7219 (2020,7).
- [Martin et al. 2020] Martin, L., Muller, B., Ortiz Suárez, P., Dupont, Y., Romary, L., Clergerie, É., Seddah, D. & Sagot, B.: "CamemBERT: a Tasty French Language Model." *Proceedings Of The 58th Annual Meeting Of The Association For Computational Linguistics*. pp. 7203-7219 (2020,7).
- [Martin et al. 2022] Martin, G., Mswahili, M., Jeong, Y. & Woo, J.: "SwahBERT: Language Model Of Swahili." *Proceedings Of The 2022 Conference Of The North American Chapter Of The Association For Computational Linguistics: Human Language Technologies*. pp. 303-313 (2022,7).
- [Micallef et al. 2022] Micallef, K., Gatt, A., Tanti, M., Plas, L. & Borg, C.: "Pre-training Data Quality and Quantity for a Low-Resource Language: New Corpus and BERT Models for Maltese." *Proceedings Of The Third Workshop On Deep Learning For Low-Resource Natural Language Processing*. pp. 90-101 (2022,7).
- [Micheli et al. 2020] Micheli, V., D'Hoffschmidt, M. & Fleuret, F.: "On the importance of pre-training data volume for compact language models." *Proceedings Of The 2020 Conference On Empirical Methods In Natural Language Processing (EMNLP)*. pp. 7853-7858 (2020,11).
- [Mikolov et al. 2013] Mikolov, T., Chen, K., Corrado, G. & Dean, J.: "Efficient Estimation of Word Representations in Vector Space." *International Conference On Learning Representations*. (2013).

- [Muller et al. 2021] Muller, B., Anastasopoulos, A., Sagot, B. & Seddah, D.: “When Being Unseen from mBERT is just the Beginning: Handling New Languages With Multilingual Language Models.” *Proceedings Of The 2021 Conference Of The North American Chapter Of The Association For Computational Linguistics: Human Language Technologies*. pp. 448-462 (2021,6).
- [Nguyen et al. 2020] Nguyen, D., Vu, T. & Tuan Nguyen, A.: “BERTweet: A pre-trained language model for English Tweets.” *Proceedings Of The 2020 Conference On Empirical Methods In Natural Language Processing: System Demonstrations*. pp. 9-14 (2020,10).
- [Nzeyimana and Rubungo 2022] Nzeyimana, A. & Rubungo, A.: “KinyaBERT: a Morphology-aware Kinyarwanda Language Model.” *ArXiv*. abs/2203.08459 (2022).
- [Ogueji et al. 2021] Ogueji, K., Zhu, Y. & Lin, J.: “Small Data? No Problem! Exploring the Viability of Pretrained Multilingual Language Models for Low-resourced Languages.” *Proceedings Of The 1st Workshop On Multilingual Representation Learning*. pp. 116-126 (2021,11).
- [Palen-Michel et al. 2022] Palen-Michel, C., Kim, J. & Lignos, C.: “Multilingual Open Text Release 1: Public Domain News in 44 Languages.” *Proceedings Of The Thirteenth Language Resources And Evaluation Conference*. pp. 2080-2089 (2022,6).
- [Parida et al. 2021] Parida, S., Biswal, S., Nayak, B., Maël, Fabien, Villatoro-Tello, E., Motlíček, P. & Dash, S.: “BERTODIA: BERT PRE-TRAINING FOR LOW RESOURCE ODIA LANGUAGE.” (2021).
- [Pennington et al. 2014] Pennington, J., Socher, R. & Manning, C.: “GloVe: Global Vectors for Word Representation.” *Proceedings Of The 2014 Conference On Empirical Methods In Natural Language Processing (EMNLP)*. pp. 1532-1543 (2014,10).
- [Peters et al. 2018] Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. & Zettlemoyer, L.: “Deep Contextualized Word Representations.” *Proceedings Of The 2018 Conference Of The North American Chapter Of The Association For Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. pp. 2227-2237 (2018,6).
- [Phang et al. 2018] Phang, J., Févry, T. & Bowman, S.: “Sentence Encoders on STILTs: Supplementary Training on Intermediate Labeled-data Tasks.” *ArXiv*. abs/1811.01088 (2018).
- [Qi et al. 2018] Qi, Y., Sachan, D., Felix, M., Padmanabhan, S. & Neubig, G.: “When and Why Are Pre-Trained Word Embeddings Useful for Neural Machine Translation?” *Proceedings Of The 2018 Conference Of The North American Chapter Of The Association For Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. pp. 529-535 (2018,6).
- [Radford and Narasimhan 2018] Radford, A. & Narasimhan, K.: “Improving Language Understanding by Generative Pre-Training.” (2018).
- [Ralethe 2020] Ralethe, S.: “Adaptation of Deep Bidirectional Transformers for Afrikaans Language.” *International Conference On Language Resources And Evaluation*. (2020).
- [Ramasesh et al. 2022] Ramasesh, V., Lewkowycz, A. & Dyer, E.: “Effect of scale on catastrophic forgetting in neural networks.” *International Conference On Learning Representations*. (2022).
- [Rocholl et al. 2021] Rocholl, J., Zayats, V., Walker, D., Murad, N., Schneider, A. & Liebling, D.: “Disfluency Detection with Unlabeled Data and Small BERT Models.” *Interspeech*. (2021).
- [Samuel et al. 2023] Samuel, D., Kutuzov, A., Øvreid, L. & Velldal, E.: “Trained on 100 million words and still in shape: BERT meets British National Corpus.” *Findings Of The Association For Computational Linguistics: EACL 2023*. pp. 1954-1974 (2023,5).
- [Schweter 2020] Schweter, S.: “BERTurk - BERT models for Turkish.” (Zenodo,2020,4).
- [Shibata et al. 1999] Shibata, Y., Kida, T., Fukamachi, S., Takeda, M., Shinohara, A. & Shinohara, T.: “Byte Pair Encoding: A Text Compression Scheme That Accelerates Pattern Matching.” (1999,9).

- [Siminyu et al. 2021] Siminyu, K., Kalipe, G., Orlic, D., Abbott, J., Marivate, V., Freshia, S., Sibal, P., Neupane, B., Adelani, D., Taylor, A., Ali, J., Degila, K., Balogoun, M., Diop, T., David, D., Fourati, C., Haddad, H. & Naski, M.: "AI4D - African Language Program." ArXiv. abs/2104.02516 (2021).
- [Souza et al. 2020] Souza, F., Nogueira, R. & Lotufo, R.: "BERTimbau: Pretrained BERT Models for Brazilian Portuguese." Brazilian Conference On Intelligent Systems. (2020,10).
- [Srivastava et al. 2014] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R.: "Dropout: A Simple Way to Prevent Neural Networks from Overfitting." J. Mach. Learn. Res. 15, 1929-1958 (2014,1).
- [Tan et al. 2018] Tan, X., Ren, Y., He, D., Qin, T., Zhao, Z. & Liu, T.: "Multilingual Neural Machine Translation with Knowledge Distillation." International Conference On Learning Representations. (2018).
- [Tela and Woubie 2020] Tela, A., Woubie, A. & Hautamäki, V.: "Transferring Monolingual Model to Low-Resource Language: The Case of Tigrinya." (2020).
- [Tsigler and Bartlett 2022] Tsigler, A. & Bartlett, P.: "Benign overfitting in ridge regression." (2022).
- [Wang et al. 2018] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O. & Bowman, S.: "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding." Proceedings Of The 2018 EMNLP Workshop BlackboxNLP: Analyzing And Interpreting Neural Networks For NLP. pp. 353-355 (2018,11).
- [Warstadt et al. 2018] Warstadt, A., Singh, A. & Bowman, S.: "Neural Network Acceptability Judgments." CoRR. abs/1805.12471 (2018).
- [Warstadt et al. 2020] Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S. & Bowman, S.: "BLiMP: The Benchmark of Linguistic Minimal Pairs for English." Transactions Of The Association For Computational Linguistics. 8 pp. 377-392 (2020).
- [Wu et al. 2016] Wu, Y., Schuster, M., Chen, Z., Le, Q., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M. & Dean, J.: "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation." (2016).
- [Wu et al. 2019] Wu, S. & Dredze, M. Beto, Bentz, Becas: "The Surprising Cross-Lingual Effectiveness of BERT." Proceedings Of The 2019 Conference On Empirical Methods In Natural Language Processing And The 9th International Joint Conference On Natural Language Processing (EMNLP-IJCNLP). pp. 833-844 (2019,11).
- [Xu et al. 2020] Xu, L., Zhang, X. & Dong, Q.: "CLUECorpus2020: A Large-scale Chinese Corpus for Pre-training Language Model." (2020).
- [Xue et al. 2020] Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A. & Raffel, C.: "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer." North American Chapter Of The Association For Computational Linguistics. (2020).
- [Zhao et al. 2021] Zhao, S., Gupta, R., Song, Y. & Zhou, D.: "Extremely Small BERT Models from Mixed-Vocabulary Training." Proceedings Of The 16th Conference Of The European Chapter Of The Association For Computational Linguistics: Main Volume. pp. 2753-2759 (2021,4).
- [Zhuang et al. 2021] Zhuang, L., Wayne, L., Ya, S. & Jun, Z.: "A Robustly Optimized BERT Pre-training Approach with Post-training." Proceedings Of The 20th Chinese National Conference On Computational Linguistics. pp. 1218-1227 (2021,8).