

Probing Language Models for Syntactic and Semantic Knowledge in African Languages

Wisdom Aduah (wizzy@aims.ac.za)
African Institute for Mathematical Sciences (AIMS)

Supervised by: Mr. Francois Meyer
University of Cape Town, South Africa

24 October 2024

Submitted in partial fulfillment of a structured masters degree at AIMS South Africa



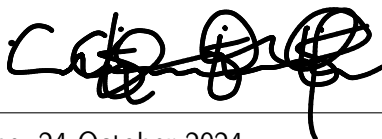
Abstract

Multilingual language models have made substantial progress in improving performance for low-resource African languages, yet the extent to which they encode linguistic knowledge for these languages remains underexplored. This study probes existing pretrained language models (PLMs) for syntactic and semantic knowledge across several African languages, using the tasks of Part-of-Speech (POS) tagging, Named Entity Recognition (NER), and News Topic Classification. I evaluate the performance of probes for different layers across several African languages, including those with limited training data like Luganda. My results indicate that while the probes achieve high performance in languages like Swahili and isiXhosa, they struggle with Igbo, suggesting the need for specialized approaches such as Multilingual Adaptive Fine-tuning (MAFT). Additionally, layer-wise analysis confirms that syntactic and semantic knowledge is mostly encoded in the middle to last layers of transformer architectures. I demonstrate the importance of adapting Pretrained Language Models on language-specific datasets, as seen in Nguni-XLMR's superior performance in isiXhosa. Using the established interpretability tool of probing, this research offers valuable insights into cross-lingual generalization and presents recommendations for future improvements in low-resource language modeling through targeted fine-tuning and transfer learning strategies.

Declaration

I, the undersigned, hereby declare that the work contained in this research project is my original work, and that any work done by others or by myself previously has been acknowledged and referenced accordingly.

Scan your signature

A handwritten signature in black ink, appearing to be 'C. D. N. N.', written over a horizontal line.

Firstname Middlename Lastname, 24 October 2024

Contents

Abstract	i
1 Introduction	1
1.1 Context and Motivation	1
1.2 Problem Statement	1
1.3 Research Questions	2
1.4 Objectives	2
1.5 Key Findings	2
1.6 Contributions	2
2 Background	4
2.1 Introduction	4
2.2 Multilingual Pretrained Language Models	4
2.3 Interpretability Through Probing Classifiers	6
2.4 Potential of Probing Classifiers	8
2.5 Probing Studies on Low-resource Languages	8
3 Methods	9
3.1 Data	9
3.2 Models	9
3.3 Probing Framework	10
3.4 Experimental Setup	13
4 Results and Discussion	14
4.1 Introduction	14
4.2 POS Tagging	14
4.3 NER labelling	16
4.4 News Topic Classification	19
5 Conclusion	22
5.1 Conclusion	22
5.2 Limitations and Future Work	22
References	28
.1 Appendix A	29

1. Introduction

1.1 Context and Motivation

The past few years have seen the proliferation of Large Language Models (LLMs) across various domains, including education, healthcare, and finance (Hadi et al., 2024). Current LLMs demonstrate impressive capabilities in modeling and understanding natural language, leading to remarkable improvement in Natural Language Processing (NLP) tasks such as translation and text generation (Millière, 2024; Li et al., 2024). However, the increasing size and complexity of these models have introduced new challenges, especially around interpretability. As a result, researchers have developed various interpretability tools—such as probing and feature attribution techniques, to gain insights into the internal workings of LLMs (Luo and Specia, 2024). Probing techniques aim to measure how well specific linguistic properties, such as part-of-speech or named entities, are encoded within the model's hidden layers. These methods are valuable because they reveal to researchers whether a model has implicitly captured linguistic knowledge during pretraining. Such insights enable a better understanding of the model's internal decision-making process, helping to identify failure modes or challenging scenarios and providing opportunities for improvement.

In parallel, there has been growing attention on Pretrained Language Models (PLMs) for African languages. Many of these languages are underrepresented in the datasets used to train existing PLMs, which limits their performance in NLP tasks. This is a pressing issue, as the absence of high-performing NLP tools for African languages restricts access to technology for millions of speakers. Efforts to address this gap have led to the development of PLMs targeting African languages, such as AfriBERTa (Ogueji et al., 2021), Afro XLMR Alabi et al. (2022), and AfroLM (Dossou et al., 2022). These models leverage strategies like transfer learning (Conneau et al., 2020) and Multilingual Adaptive Fine-tuning (MAFT) (Alabi et al., 2022) to improve performance on low-resource languages.

Despite the progress in building PLMs for African languages, there is limited understanding of how well these models encode fundamental linguistic properties for these languages. While probing has been applied to PLMs trained on high-resource languages like English and French, to the best of my knowledge, it has not yet been used to study linguistic understanding in African languages. Understanding which linguistic features are effectively encoded in these models can guide the fine-tuning process for downstream tasks. For example, knowing which layers capture syntactic properties like part-of-speech or semantic features like named entities can inform where to apply fine-tuning efforts, optimizing both performance and efficiency (Katinskaia and Yangarber, 2024).

1.2 Problem Statement

1.2.1 Lack of Research for African Languages

While transformer-based PLMs have been widely studied for high-resource languages like English, French, and Chinese, there is still limited research on probing these models for African languages. Many African languages are underrepresented in the training data, and it is not clear whether these models encode essential linguistic properties, such as syntactic knowledge (e.g., part-of-speech) and semantic knowledge (e.g., named entities) for these languages.

1.2.2 Consequences of the Gap in Research

If a model poorly encodes linguistic information for these languages, this will likely affect downstream tasks. For instance, if a model struggles to distinguish named entities in Swahili, it will not perform well on Named Entity Recognition (NER) tasks for that language. This directly affects their real-world applications. For African languages, models may be employed for critical tasks like machine translation, information extraction, or automatic text classification. A model that does not encode the language well is less reliable for these applications, which could further magnify the digital divide for speakers of low-resource languages.

1.3 Research Questions

My research questions are as follows:

1. To what extent do the hidden layers of PLMs encode knowledge about grammatical categories and semantic topics in African languages?
2. Which specific layers in the transformer-based PLMs contain the most information relevant to part-of-speech tagging, named entity recognition, and news topic classification?
3. How does probe performance vary across African languages and relate to the level of language resourcedness?

1.4 Objectives

The goal of this study is to design and train probe classifiers on the internal representations of seven PLMs to reveal insights into how they encode syntax and semantics across six African languages. This work aims to pioneer the training of probes specifically for African languages, by applying established strategies for evaluating probe performance in the context of these languages. Finally, the study focuses on analyzing variations across different models, languages, and tasks to better understand their behavior and performance.

1.5 Key Findings

My key findings reveal that:

1. Adapting pretrained language models to a few languages improves probing performance.
2. POS and NER knowledge is mostly prevalent in middle-to-last hidden layers, while for NTC, the knowledge is almost evenly distributed across layers.
3. Multilingual PLMs demonstrate cross-lingual transfer ability to low-resource languages from their closely related high-resource counterparts.

1.6 Contributions

My contributions are as follows:

1. I designed probes for analyzing Afro XLMR, XLM-R, Nguni-XLMR, AfriBERTa, and AfroLM on their linguistic abilities on Part-of-Speech, Named Entity Recognition, and News Topic Classification tasks in 6 African languages.
2. I designed a control task for MasakhaPOS, and trained random baseline probes for MasakhaNER and MasakhaNEWS datasets.
3. I analyzed the role of multilingual training strategy and language resourcedness in probe performance.

2. Background

2.1 Introduction

This chapter opens with a review of selected masked transformer (Vaswani et al., 2017) multi-lingual PLMs, which my work is focused on. It then discusses probing classifiers, their limitations, proposed solutions from the literature, and their potential. Lastly, I provide a brief overview of probing studies on both high- and low-resource languages.

2.2 Multilingual Pretrained Language Models

Multilingual PLMs have become foundational tools in natural language processing (NLP), with transformer-based models being the most popular. These models, however, are often trained on unevenly distributed language resources, favoring high-resource languages over low-resource ones. In this work, I focus on the languages Swahili, isiXhosa, Naija Pidgin, Hausa, Igbo, and Luganda. The following is an overview of the key PLMs I used in this study. They represent a broad set of approaches to developing multilingual PLMs for African languages, from massively multilingual modelling to narrowing the linguistic scope, as well as adapting PLMs for African languages.

2.2.1 Cross-lingual Language Model - RoBERTa (XLM-R)

XLM-R is a cross-lingual, encoder-only transformer-based model, derived from RoBERTa (Conneau et al., 2020). Pretrained on approximately 2.5 terabytes of multilingual data from the Common Crawl corpus, it covers 100 languages, including 8 African languages (representing just 0.2% of the dataset). Out of these, Hausa, Swahili, and isiXhosa are utilized in this study.

One of XLM-R’s important contributions is cross-lingual transfer, where a model trained on high-resource languages can generalize well to low-resource ones. The authors found that increasing language coverage while maintaining model capacity improves transfer performance, up to a certain point. XLM-R surpassed mBERT on several benchmarks, including cross-lingual natural language inference (XNLI), Question Answering, and Named Entity Recognition (NER), with notable improvements on low-resource languages.

However, despite these improvements, the gap in performance between low and high-resource languages in XLM-R remains significant as Conneau et al. (2020) show in their experiments. Furthermore, increasing language coverage requires vast amounts of data, resulting in large model sizes. There are several pretrained versions, but my study is focused on XLM-R Base (270M parameters, 250K vocab size, 12 layers) and XLM-R Large (550M parameters, 250K vocab size, 24 layers).

2.2.2 AfroXLMR

AfroXLMR builds upon XLM-R and is adapted specifically for African languages through a technique called Multilingual Adaptive Fine-Tuning (MAFT) (Alabi et al., 2022). AfroXLMR covers 17 African languages along with three high-resource languages (English, Arabic, and French). It has achieved state-of-the-art performance on tasks such as NER, sentiment classification, and news topic classification for African languages.

AfroXLMR-Base and AfroXLMR-Large, are adapted versions of XLM-R Base and Large, respectively. A

lightweight version, AfroXLMR-Small, reduces the vocabulary size to 70K tokens, halving the model size with minimal performance degradation. Despite not including Luganda in the training data, AfroXLMR performed strongly across African languages, reducing the performance gap between high- and low-resource languages.

2.2.3 AfriBERTa

Instead of leveraging on the knowledge of existing pretrained models, Ogueji et al. (2021) pretrained three relatively small transformer models from scratch on 11 African languages using less than 1GB of training data. Their aim was to show that pretraining on a small dataset could produce results comparable to the benefits of cross-lingual transfer from high-resource languages. The three models are detailed in Table 2.1.

Model	Parameters	Vocab size	Layers	Model size
AfriBERTa large	126M	70K	10	768
AfriBERTa base	111M	70K	8	768
AfriBERTa small	97M	70K	4	768

Table 2.1: AfriBERTa pretrained models

They fine-tuned their models on Named Entity Recognition (NER), and text classification downstream tasks using datasets from Adelani et al. (2022) and Hedderich et al. (2020) respectively. In NER, their models outperform both mBERT and XLM-R in Amharic, Hausa, Igbo, Swahili and Yoruba. They were outperformed in 5 languages. Out of these 5, three of them were not represented in the training data of any of the models. The other 2 languages were represented in the training dataset of AfriBERTa but not in mBERT or XLM-R. This showed the advantage of cross-lingual transfer enabled by larger models trained on more languages.

For text classification, they used a dataset with only two languages (Hausa and Yoruba). AfriBERTa outperformed mBERT and XLM-R in this task. However, it is hard to draw conclusions since only 2 languages were evaluated. However, considering how significantly smaller AfriBERTa models are, these results established the feasibility of smaller-scale PLMs to improve performance for specific languages.

2.2.4 AfroML

Inspired by the work on AfriBERTa, Dossou et al. (2022) aimed to train a language model from scratch using an even smaller dataset covering 23 African languages. However, their model featured a significantly larger vocabulary (250K) and more parameters (264M). Unlike traditional unsupervised learning approaches, they employed self-active learning for training.

Their evaluation baselines included mBERT, XLM-R, AfroXLMR-base, and another variant of AfroLM trained from scratch without active learning. The models were fine-tuned on Named Entity Recognition (NER), text classification, and sentiment analysis tasks.

In NER, AfroLM with active learning outperformed AfriBERTa in 7 out of 10 languages. It also outperformed AfroXLMR-base in only two languages (Ibo and Luo) but underperformed in the remaining 8. Similarly, it outperformed mBERT and XLM-R base in 6 languages but was outperformed in Kinyarwanda, Luo, Luganda, and Wolof—languages included in AfroLM’s pretraining data but absent from mBERT and XLM-R base.

For text classification, only AfriBERTa and AfroLM without active learning served as baselines, with evaluations limited to Hausa and Yoruba. AfroLM with active learning showed slight improvement over AfriBERTa in Hausa and outperformed its non-active-learning variant in both languages. However, AfriBERTa outperformed AfroLM in Yoruba.

Considering that XLM-R base, mBERT, and AfroXLMR-base were pretrained on datasets 14 times larger than that used for AfroLM, these results are impressive and highlight the potential of AfroLM's approach.

2.2.5 Nguni-XLMR

Similar to Afro XLMR, Nguni-XLMR was created by applying Multilingual Adaptive Fine-tuning to XLM-R-large on 4 Nguni languages (isiXhosa, isiZulu, isiNdebele, and Siswati) (Meyer et al., 2024). Their goal was to address the problem of the *curse of multilinguality* by focusing on a smaller number of low-resource languages. Using Afro XLMR-large and XLM-R-large as baselines, their results showed significant performance improvements across 4 tasks (Named Entity Recognition, Part-of-Speech tagging, News Topic Classification, and Phrase Chunking). The curse of multilinguality refers to the trade-off between maintaining cross-lingual performance on low-resource languages and expanding language coverage, given a fixed model capacity. Beyond a certain threshold of language coverage, performance on individual languages tends to decline.

2.3 Interpretability Through Probing Classifiers

PLMs generate rich, contextualized word representations of input texts, which are essential to their success across a wide range of NLP tasks. One way to analyse and interpret these representations is through probing, which investigates how much linguistic information the model encodes.

Structural probing tries to find correlations between the internal representations of the model and some linguistic property of interest. This is usually done by training a classifier (probe) to extract information from the contextualized representation. If the probe performs well in this task, then the language model is said to encode relevant information about the linguistic property. One of the earliest form of this technique analysed uncontextualized word embeddings for linguistic properties using a classifier (Köhn, 2015).

Though independent of model architecture, probing classifiers rely heavily on the availability of annotated datasets. Belinkov (2022) noted that the choice of which linguistic property to probe for is often constrained by the availability of annotated data for that task. This situation is undesirable for low resource languages, as there is little annotated datasets targeting these properties. Moreover, the annotated data may miss on other features, important to the model (Michael et al., 2020).

One of the challenges in probing is interpreting the results. For instance, if a probe achieves an accuracy of 80%, it is not clear whether this should be considered high or low without a suitable baseline, as the probe could have simply learned the task using heuristics unrelated to the probing task itself, without actually relying on the knowledge encoded in PLM representations. The following are some strategies that have been proposed to contextualize probe performance.

2.3.1 Baselines

Researchers have used baselines such as majority class prediction (Belinkov et al., 2017; Conneau et al., 2018b) or random hidden representations (Zhang and Bowman, 2018b; Conneau et al., 2018b; Chrupała

et al., 2020; Tenney et al., 2019) to contextualize probe performance. However, as noted by Hewitt and Liang (2019), random baselines might still encode information, which a sophisticated classifier could exploit.

2.3.2 Control Tasks

A major concern in probing is whether a probe is learning the linguistic property of interest or simply memorizing the task. To address this, control tasks have been proposed (Hewitt and Liang, 2019). These tasks are designed to test the model's ability to memorize information, allowing researchers to distinguish between learning the task and encoding useful linguistic information. They formalize this distinction (which they call selectivity), as probing performance on original task minus performance on control task. Nonetheless, Pimentel et al. (2020) argue that memorization is an essential component of linguistic competence and that the distinction between task learning, and property encoding may be artificial. Moreover, control tasks by design are limited to word-level tasks.

2.3.3 Control functions

Another approach proposed for addressing probe interpretability involves measuring information gain by applying a function known as a control function to the internal representations (Pimentel et al., 2020). This approach, unlike control tasks, is not limited to word-level tasks and offers a more generalizable framework for probing.

2.3.4 Control Datasets

Ravichander et al. (2020) suggests control datasets for contextualizing probing tasks. They modify the original pretraining dataset, to only have one value for the linguistic property in all samples. The intuition is that, any model trained on this data will have no knowledge of the linguistic property — hence a probe's ability to still perform well on such a task is evidence that the probe is not learning any useful information about the linguistic property of interest. They confirm from their experiments that a probe may indeed still do well on a model that has not learned the task.

2.3.5 Restrictions on Classifier

Some studies advocate various restrictions on the probe classifier in terms of its complexity, and size of the training data. While their focus was on image models, Alain and Bengio (2018) argue that the objective of a deep neural network classifier is to produce in its final layer, a representation that can easily be passed to a linear classifier. In other words, a linear probe is the natural choice for probing. They also state the benefit of convergence to the global minima when a linear classifier is trained with a Softmax Cross-entropy — which effectively avoids the problem of local minima.

Following Alain and Bengio (2018), Hewitt and Liang (2019) went a step further by proposing constraints on the probe and argue that more expressive probes are prone to memorization (low selectivity), and advocate for linear probes. They determined that constraining the number of training examples (400 out of 39832), weight decay constant (0.1 to 0.01), and hidden state dimensionality (10 to 50) improved *selectivity*.

On the other hand, Conneau et al. (2018b) suggests that some linguistic features might not be linearly separable in the representation, while Pimentel et al. (2020) argue for complex probes for retrieving information about a linguistic property.

2.4 Potential of Probing Classifiers

Despite advances in the probing classifier framework, there is still no reliable way of accurately interpreting their outcomes. However, they are easy to implement and are independent of model architecture. Unlike downstream tasks, which aim to achieve high performance on practical, real-world tasks such as machine translation, probing tasks investigate the internal black box representations of a model to reveal fundamental linguistic properties encoded within those representations.

2.5 Probing Studies on Low-resource Languages

Numerous probing research target high resourced languages like English, French, and Russian (Arps et al., 2024; Zhang and Bowman, 2018a; Hewitt and Liang, 2019; Katinskaia and Yangarber, 2024; Conneau et al., 2018a; Hou et al., 2024), probably due to availability of annotated datasets. While others (Arora et al., 2022; Li et al., 2024) have explored some low resource languages like Hindi, and Tamil — to the best of my knowledge, there has not been any research targeting African languages.

3. Methods

3.1 Data

Data for my experiments were obtained from MasakhaNER (Adelani et al., 2022) for Named Entity Recognition (NER), MasakhaPOS (Dione et al., 2023) for Part-of-Speech (POS) labeling, and MasakhaNews (Adelani et al., 2023) for News Topic Classification (NTC).

The NER dataset is annotated with *PER*, *ORG*, *LOC*, and *DATE* named entities, covering 20 African languages. This dataset is the most extensive public resource for NER tagging in African languages.

For the POS dataset, it is annotated with 17 POS tags and also covers 20 African languages. The full list of labels for the POS tags can be found at UD¹.

The NTC dataset is annotated with 7 news topic labels (business, entertainment, health, politics, religion, sports, technology), and covers 16 African languages.

Table 3.1 summarizes the 3 datasets for the 6 chosen languages for this work. The splits (train, validation, and test) are stated in terms of number of sentences (number of news articles for NTC).

The languages were selected based on the availability of datasets for all three tasks, while also ensuring linguistic diversity. Additionally, some languages were intentionally chosen to be absent from the pre-training data of certain models, allowing us to evaluate how well these models perform on languages they had not seen during pretraining or adaptation (See Table 3.2).

Language	Family	Region	Train			Validation			Test		
			NER	POS	NTC	NER	POS	NTC	NER	POS	NTC
Hausa	Afro-Asiatic	West	5716	753	2219	816	150	317	1633	601	637
Igbo	Niger-Congo	West	7634	803	1356	1090	160	194	2181	642	390
Luganda	Niger-Congo	East	4942	733	771	706	146	110	1412	586	223
Naija Pidgin	Creole	West	5646	586	1060	806	150	152	1294	600	305
Swahili	Niger-Congo	East	6593	693	1658	942	138	237	1883	553	476
isiXhosa	Niger-Congo	Southern	5718	752	1032	817	150	147	1633	601	297

Table 3.1: Summary of NER, POS, and NTC datasets across training, validation, and test sets.

3.2 Models

Seven pretrained language models were selected for this study, guided by the study’s objectives. The selection included models specifically pretrained on African languages (AfriBERTa, and AfroLM), those adapted for African languages (Afro XLMR-large, Afro XLMR-base, and Nguni-XLMR), and others in which African languages were significantly underrepresented (XLM-R models). Details of the models are discussed in section 2.2. Details of their configurations are shown in Table 3.3

¹<https://universaldependencies.org/u/pos/>

Model	Swahili	Igbo	Hausa	Luganda	isiXhosa	Naija Pidgin
Afro XLMR-large	★★	★★	★★	☆☆	★★	★★
Afro XLMR-base	★★	★★	★★	☆☆	★★	★★
AfriBERTa	★★	★★	★★	☆☆	☆☆	★★
AfroLM	★★	★★	★★	★★	★★	★★
XLM-R-large	★★	☆☆	★★	☆☆	★★	☆☆
XLM-R-base	★★	☆☆	★★	☆☆	★★	☆☆
Nguni-XLMR	★☆☆	★☆☆	★☆☆	☆☆	★★	☆☆

Table 3.2: Language coverage of pretrained models. 0 star - no data from the language was included either in pretraining or adaptation. 1 star: The language was included in the base model but not in the adapted model. 2 stars: The model was either pretrained or adapted for the language.

Model	Parameters	Vocab size	Layers	Model size	Max token input length
Afro XLMR-large	550M	250K	24	1024	512
Afro XLMR-base	270M	250K	12	768	512
AfriBERTa large	126M	70K	10	768	512
AfroLM	264M	250K	10	768	256
XLM-R large	550M	250K	4	1024	512
XLM-R base	270M	250K	12	768	512
Nguni XLM-R	550M	250K	24	1024	512

Table 3.3: Configurations of pretrained models

3.3 Probing Framework

To assess the model’s ability to encode syntactic, semantic, and contextual information, I conducted three probing tasks, each targeting different linguistic properties:

1. Named Entity Recognition
2. Part-of-Speech Tagging
3. News Topic Classification

I used a Multilayer Perceptron (MLP) with 1 hidden layer for all three tasks. Table 3.4 summarizes the MLP architecture.

Hyperparameter	POS	NER	NTC
Activation function	ReLu	ReLu	ReLu
Loss function	Cross entropy	Cross entropy	Cross entropy
Hidden layers	1	1	1
Hidden state dimensionality	50	50	50
Learning rate	.001	.001	.00001
Batch size	32	32	2
Optimizer	Adam	Adam	Adam
Dropout	.2	.2	.2

Table 3.4: Probe classifier hyperparameters

Formally, the MLP is defined as follows:

$$\mathbf{y} = f(\mathbf{W}_2 \sigma(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2)$$

, where $\mathbf{x} \in \mathbb{R}^n$ is the input representation vector for a word, $\mathbf{W}_1 \in \mathbb{R}^{m \times n}$ is the weight matrix for the hidden layer, $\mathbf{W}_2 \in \mathbb{R}^{k \times m}$ is the weight matrix for the output layer, $\mathbf{b}_1 \in \mathbb{R}^m$ is a bias vector, $\sigma(\cdot)$ is the ReLU activation function, and $f(\cdot)$ is a softmax function.

3.3.1 Word-level Tasks

For NER and POS tasks, I define a word-level task as a function f that maps an input sequence X , to an output sequence Y . That is $f : X \rightarrow Y$, where X is the sequence of contextualized word embeddings of the input text, and Y is the sequence of output labels for each word embedding in X .

For example, given the input sequence [He, worked, alone], the pretrained model generates contextualized embeddings for each subword in each layer. The probe then predicts the label for each subword embedding for a specific layer as illustrated in Figure 3.1.

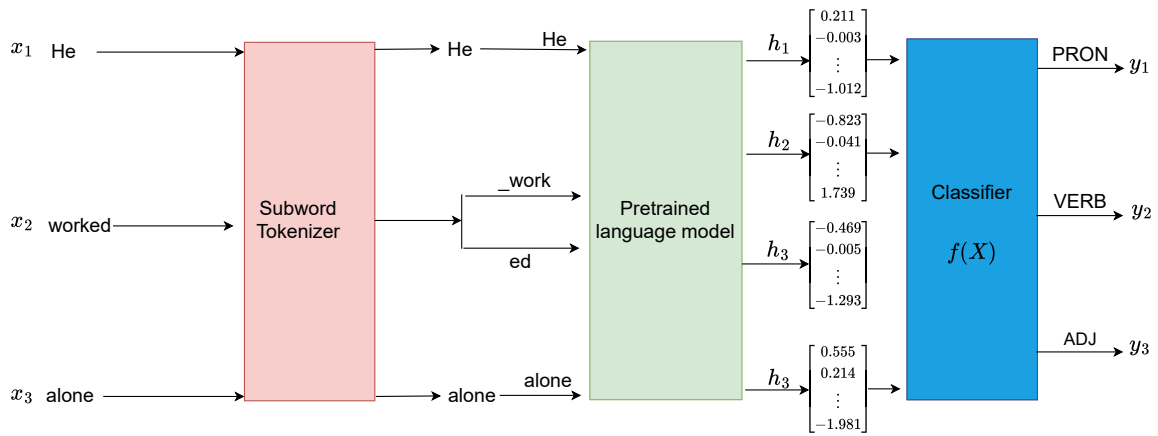


Figure 3.1: Probing task illustration. h_1 , h_2 , and h_3 are the hidden state representation in an arbitrary layer. 'worked' is tokenized into '_work' and 'ed', but the representation of '_work' (first subword pooling) is chosen to be passed to the classifier.

Given that some words are tokenized into multiple subwords, I used only the first subword representation as input for the classifier to align the words with their hidden representations. Alternative strategies include using the last subword, averaging the subword embeddings, or applying attention over the subwords—approaches I was unable to explore due to time constraints. Ács et al. (2021) demonstrates that the choice of subword pooling can make a difference in performance, especially in languages with high subword counts.

POS tagging was evaluated using accuracy, while the F1-score was used for NER evaluation. For NER, non-entity labels (O) accounted for over 90% of the dataset, making F1-score a more appropriate metric.

3.3.2 Sentence-level Tasks

A sentence-level task for News Topic Classification is defined similarly as the word-level one, except that the input to the classifier is not individual word embeddings of words. Here, only the word embeddings for the special classification token (<s> token for XLM-R base models) are passed as input to the classifier, since they represent the entire input sequence. For input texts longer than the maximum input token length for the model, the texts were truncated to fit the desired input size.

3.3.3 Control Tasks

In order to give context to the POS probing results, I used control tasks (Hewitt and Liang, 2019). The design of control tasks is achieved in two major steps:

1. Definition of random control behavior for each word type
2. Deterministic assignment of labels to words in the dataset

For each word w_i in the vocabulary V of the language L , a random label is assigned to w_i (control behavior). This control behavior is then used to deterministically re-assign labels to each word in the dataset (train, dev, and test) to produce a control dataset. The control dataset is then used for the probing task (Control task). It is important to note the following:

1. Samples are drawn from the empirical distribution of the labels.
2. By design, control tasks are originally not applicable to sentence-level tasks like topic classification.
3. Control tasks are designed to have both structure and randomness.

Control tasks are designed to solve a task by memorizing input-output pairs, allowing the performance of the probe on that task to be put into context. Hewitt and Liang (2019) define **Selectivity** as:

$$\text{selectivity} = \text{accuracy on probing task} - \text{accuracy on control task}$$

Thus, selectivity tells us what a probe can achieve without memorization.

Although NER is a word-level task, applying control tasks to it was challenging for the following reasons:

1. It was not clear whether control tasks should assign labels to individual words or entire entity spans.
2. Both word-level and span-level control tasks resulted in negative selectivity, making the interpretation difficult.

3.3.4 Random Baselines

As mentioned in section 3.3.3, control tasks in their original form are not applicable to sentence-level tasks, and challenging to implement for NER. Therefore, I included the following random baselines:

1. Randomly initialized BiLSTM for NTC
2. Randomly initialized transformer model for NER

The input to the untrained BiLSTM consists of uncontextualized word embeddings generated by a character-level Convolutional Neural Network (CNN). These embeddings are then contextualized using a randomly initialized BiLSTM layer with a hidden size of 1024. This setup serves as a strong baseline, enabling us to assess whether the contextualized representations of the pretrained models provide any

advantage over purely random contextualizations (Conneau et al., 2018b; Hewitt and Manning, 2019). Output of the LSTM was aggregated using mean pooling to obtain a single representation for the input text. I used this baseline for News topic classification.

I randomly initialized a transformer model (AfriBERTa) and used its first contextualized layer as baseline for all the models in NER task. The selection of AfriBERTa was arbitrary. Additionally, I constructed a separate baseline by randomly initializing all six transformer models to serve as a second transformer baseline (Zhang and Bowman, 2018b; Conneau et al., 2018b; Chrupała et al., 2020; Tenney et al., 2019). In this second setup, each pretrained model has a corresponding randomly initialized version where layerwise comparison could be made. I then computed the **gain** as:

$$\text{gain} = \text{f1-score on original model} - \text{f1-score on randomly initialized model}$$

3.4 Experimental Setup

3.4.1 Hardware Resources

I used a VM instance with the following configuration:

1. 1 NVIDIA T4 GPU with 15 GB memory.
2. 2 virtual CPU cores with 20 GB RAM.
3. 1 TB SSD Disk space.

3.4.2 Coding Environment

All my coding was done in Google Colab. For POS and NER, the source code was adapted from ². The adapted code and NTC code I wrote are available as well. ³

3.4.3 Libraries and Frameworks

I used the following libraries and frameworks:

1. Seqeval to evaluate Named Entity Recognition tasks ⁴.
2. Huggingface's transformer library ⁵ for loading and working with the pretrained models.
3. Huggingface's datasets ⁶ library for loading some of my datasets.
4. PyTorch for loading datasets and training my MLP classifiers.

3.4.4 Runtime

Using the hyperparameter configurations from Table 3.4: Maximum running time for NER task per language was 13 hours. Maximum running time for POS task per language was 10 hours. Maximum running time for NTC task per language was 12 minutes

²<https://github.com/juditacs/probing/blob/main/README.md>

³<https://github.com/Anaphase21/Probing-language-models-for-syntactic-and-semantic-knowledge-on-African-languages>

⁴<https://pypi.org/project/seqeval/>

⁵<https://github.com/huggingface/transformers>

⁶<https://github.com/huggingface/datasets>

4. Results and Discussion

4.1 Introduction

I present the results and discussion in three parts: POS tagging, NER labeling, and NTC. Raw accuracies, F1-scores, and individual layer results are omitted here but can be found in Appendix A. The tables also include *Coverage* ratings, which assess the model’s language support. 0 star: the language was not seen both in the base model and the adapted model. 1 star: The language was included in the base model but not in the adapted model. 2 stars: The model was either pretrained or adapted for the language.

4.2 POS Tagging

Language-wise, all seven models show their weakest performance in Igbo, both in terms of raw accuracy as shown in Figure 4.1 and selectivity. On the other end, they achieve their highest accuracies in Swahili and Hausa. Luganda records the second-highest accuracy and selectivity, despite not being represented in all but AfroLM. This outcome suggests strong cross-lingual generalization to Luganda.

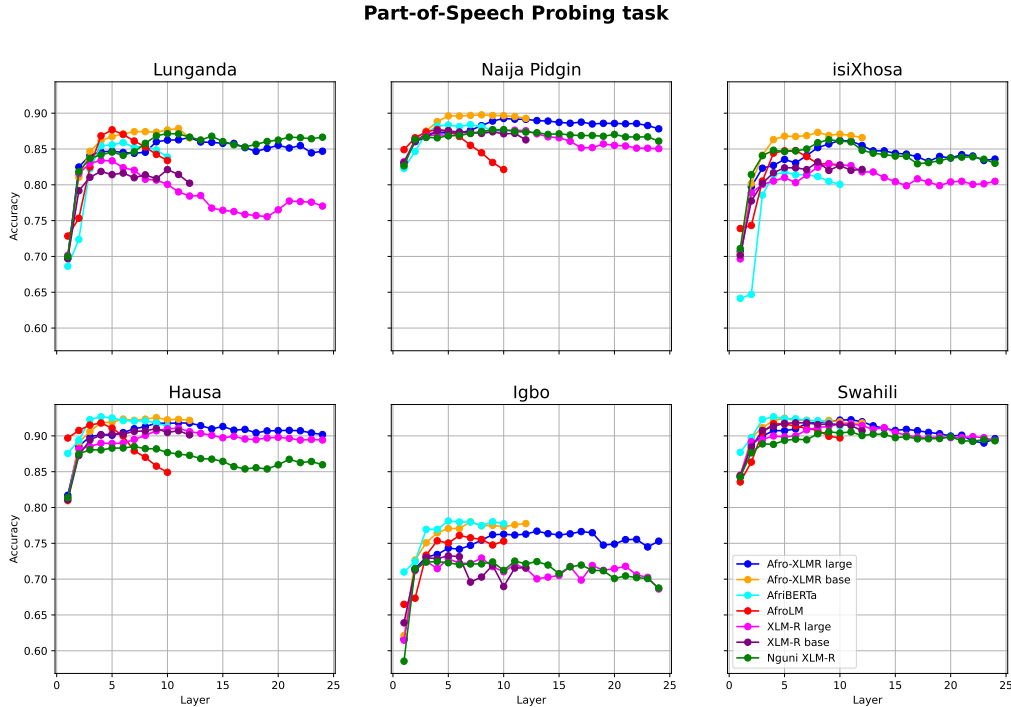


Figure 4.1: POS accuracy across all 6 layers and languages.

To better contextualize the raw accuracies, I analyze the selectivity scores shown in Figure 4.2 and Table 4.1. Selectivity helps show the degree to which each model captures part-of-speech information in its internal layers as already discussed in the methods section. All models achieve positive selectivity across all layers, except in Igbo, indicating that POS features are generally encoded in their representations. As expected, Nguni XLM-R—fine-tuned specifically on isiXhosa—achieves the highest selectivity (43.13%) in isiXhosa, agreeing with its specialization for this language. In contrast, Igbo presents the

lowest selectivity across models, confirming their struggle to encode POS information effectively in this language.

Among the models, AfroLM stands out by recording the highest selectivity in four of the six languages: Naija Pidgin, Luganda, Hausa, and Swahili. AfriBERTa, despite being the smallest model, is competitive with AfroXLMR-large in Hausa and Swahili. Interestingly, AfroXLMR-large performs best in Igbo with a selectivity score of 7.62%, although this score is still relatively low compared to the other languages.

The models also demonstrate cross-lingual transfer abilities as shown in Table 4.1. For example, even though Luganda is not covered in AfroXLMR-large, the model achieves a selectivity score close to that of AfroLM, exhibiting its ability to transfer POS knowledge to previously unseen languages. This suggests some capacity for cross-lingual transfer, especially from other closely related languages (Bantu languages such as Swahili and isiXhosa), which aligns with the idea that these models can generalize semantic knowledge to unseen or underrepresented languages.

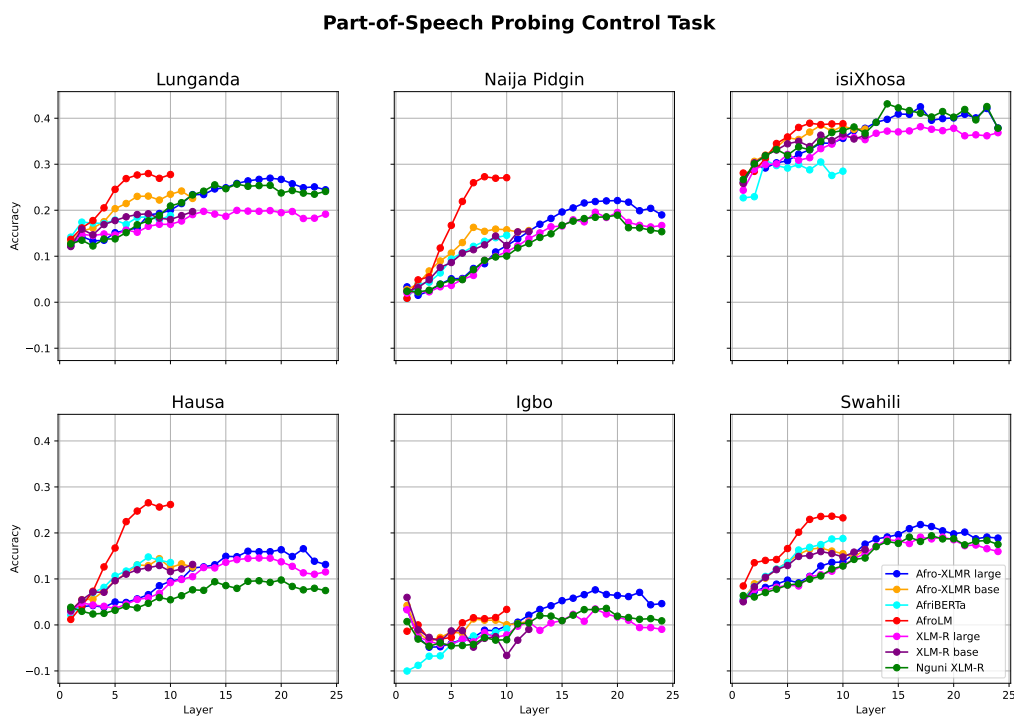


Figure 4.2: POS Selectivity across all layers and languages.

Finally, a clear pattern emerges in how POS information is encoded across the transformer layers. In general, the early layers capture less POS knowledge compared to later ones, with the trend peaking or plateauing around the middle layers. This pattern aligns with existing literature on transformers, where mid-to-last-layers often encode syntactic features more effectively (Li et al., 2024).

		swa	ibo	hau	lug	xho	pcm
Afro XLMR Large	Best layer	17	18	22	19	17	19
	Selectivity	21.82	7.62	16.56	26.97	42.49	22.03
	Coverage	★★	★★	★★	☆☆	★★	★★
Afro XLMR Base	Best layer	8	1	9	11	8	7
	Selectivity	16.73	4.22	14.40	24.18	38.48	16.28
	Coverage	★★	★★	★★	☆☆	★★	★★
AfriBERTa	Best layer	10	10	8	10	3	10
	Selectivity	18.79	-0.82	14.77	18.85	30.87	14.55
	Coverage	★★	★★	★★	☆☆	☆☆	★★
AfroLM	Best layer	9	10	8	8	7	8
	Selectivity	23.64	3.37	26.55	27.98	38.92	27.28
	Coverage	★★	★★	★★	★★	★★	★★
XLM-R Large	Best layer	17, 19	1	18	16	17	18
	Selectivity	19.09	3.29	14.54	19.97	38.15	19.56
	Coverage	★★	☆☆	★★	☆☆	★★	☆☆
XLM-R Base	Best layer	12	1	12	8	8	12
	Selectivity	16.39	5.98	13.15	19.23	36.32	15.39
	Coverage	★★	☆☆	★★	☆☆	★★	☆☆
Nguni-XLMR	Best layer	16	19	20	16	14	20
	Selectivity	19.09	3.6	9.76	25.67	43.13	18.92
	Coverage	★☆☆	★☆☆	★☆☆	☆☆	★★	☆☆

Table 4.1: Selectivity for POS across 6 languages. The highest selectivity for each language is boldened. 'Coverage' rates the extent of coverage for the language. 0 star: the language was not seen neither in base model or the adapted model. 1 star: The language was included in the base model but not in the adapted model. 2 stars: The model was either pretrained or adapted for the language.. Full table for layers is in Appendix A.

4.3 NER labelling

The results from the NER task across the six languages indicate similar general trends to those observed in the POS task, with performance starting low in the early layers and peaking around the middle layers. Instead of using a control task as a baseline, I used the first contextualized layer of the transformer model, represented by the black line in Figure 4.3, to provide a point of comparison. As discussed in the methods section, this was done because I could not find a way to design control tasks for NER.

All models exceed the baseline F1-score across all six languages, confirming that they encode meaningful information relevant to NER. However, the extent of this encoding varies between languages, providing insights about one of my research questions on the models' ability to capture and generalize semantic knowledge across diverse, low-resource languages.

Using a layer-wise random baseline for each model as shown in Figure 4.4 and Table 4.2, confirms Igbo to be indeed the lowest performing language. Although Luganda is not represented in all but AfroLM,

Afro XLMR-large and Nguni XLM-R still produced very high gains. This suggests some capacity for cross-lingual transfer, particularly from other closely related languages (Bantu languages such as Swahili and isiXhosa), which aligns with the idea that these models can generalize semantic knowledge to unseen or underrepresented languages. Nguni XLM-R, as expected, achieves a high F1-score (82.23%) and gain (58.46%) in isiXhosa due to its fine-tuning on the Nguni language family, which shows the importance of targeted fine-tuning for improved model performance.

The highest raw F1-scores are recorded in Swahili (88.94%) and Igbo (82.58%), with Swahili performing exceptionally well across most models. However, the gap between the model performances and the baseline is narrower compared to other languages. Using layer-wise baselines as shown in Figure 4.4, the performances on Igbo is seen to be poor compared to other languages.

		swa	ibo	hau	lug	xho	pcm
Afro XLMR Large	Best layer	22	17	22	16	19	17
	Gain	78.41	55.20	70.63	75.79	77.96	76.27
	Coverage	★★	★★	★★	☆☆	★★	★★
Afro XLMR base	Best layer	12	8	8	12	8	8
	Gain	47.07	50.55	66.06	57.7	65.27	69.52
	Coverage	★★	★★	★★	☆☆	★★	★★
AfriBERTa	Best layer	9	10	5	7	10	8
	Gain	42.45	53.78	64.52	48.75	67.36	57.80
	Coverage	★★	★★	★★	☆☆	☆☆	★★
AfroLM	Best layer	6	9	4	8	8	6
	Gain	39.45	46.33	61.21	45.90	62.07	47.97
	Coverage	★★	★★	★★	★★	★★	★★
XLM-R large	Best layer	23	16	19	20	20	13
	Gain	79.98	45.03	66.63	66.95	68.97	74.20
	Coverage	★★	☆☆	★★	☆☆	★★	☆☆
XLM-R base	Best layer	11	10	7	12	10	12
	Gain	41.54	37.79	60.84	60.95	55.16	58.46
	Coverage	★★	☆☆	★★	☆☆	★★	☆☆
Nguni-XLMR	Best layer	22	22	20	21	20	21
	Gain	73.10	47.52	60.57	77.28	80.18	58.46
	Coverage	★☆☆	★☆☆	★☆☆	☆☆	★★	☆☆

Table 4.2: Gain for NER across 6 languages. Highest gain for each language is boldened. 'Coverage' rates the extent of coverage for the language. 0 star: the language was not seen neither in base model or the adapted model. 1 star: The language was included in the base model but not in the adapted model. 2 stars: The model was either pretrained or adapted for the language. Full table for layers is in appendix A.

On the other end, the models' performance is less impressive in Hausa, based on the raw F1-scores alone, with most models scoring around 75%. However, the baseline is far lower than in other languages, giving it better performance than for example Igbo.

The most interesting result is the wide gap between model performance and baseline in isiXhosa, which reveals that the models encode a substantial amount of NER-specific knowledge for this language. This also emphasizes the role of MAFT, as Nguni XLM-R, which was specifically adapted for isiXhosa, achieves the best performance among models tested on this language. However, the relatively lower performance in Igbo compared to other languages reflects the challenge of encoding NER features in underrepresented languages, despite Igbo being present in the pre-training data.

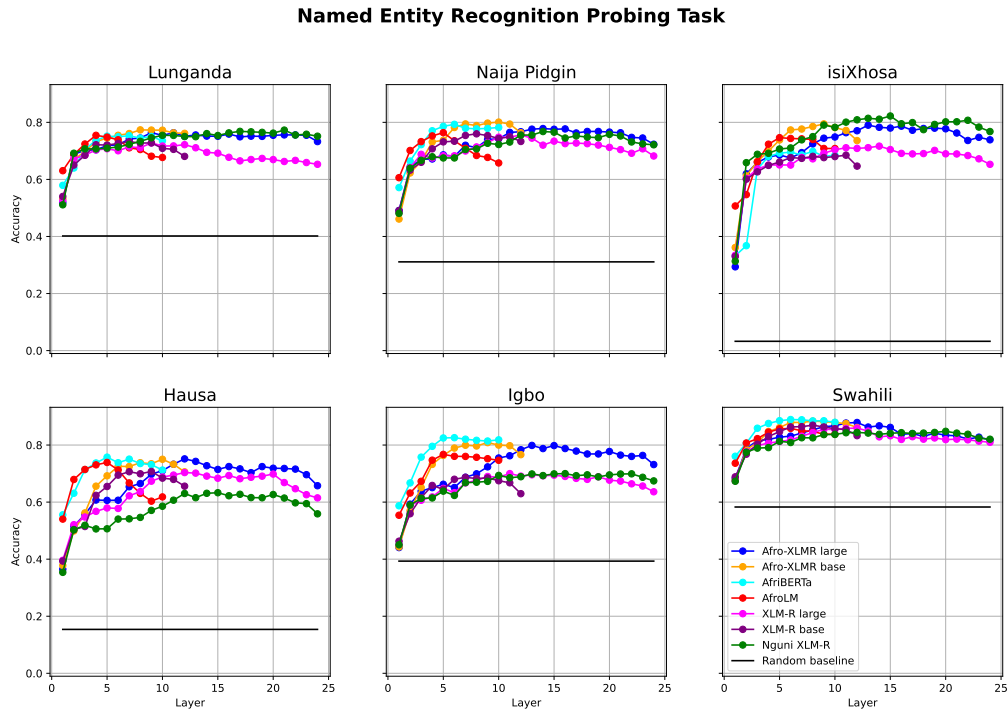


Figure 4.3: F1-score for NER task. The black line represents the random baseline

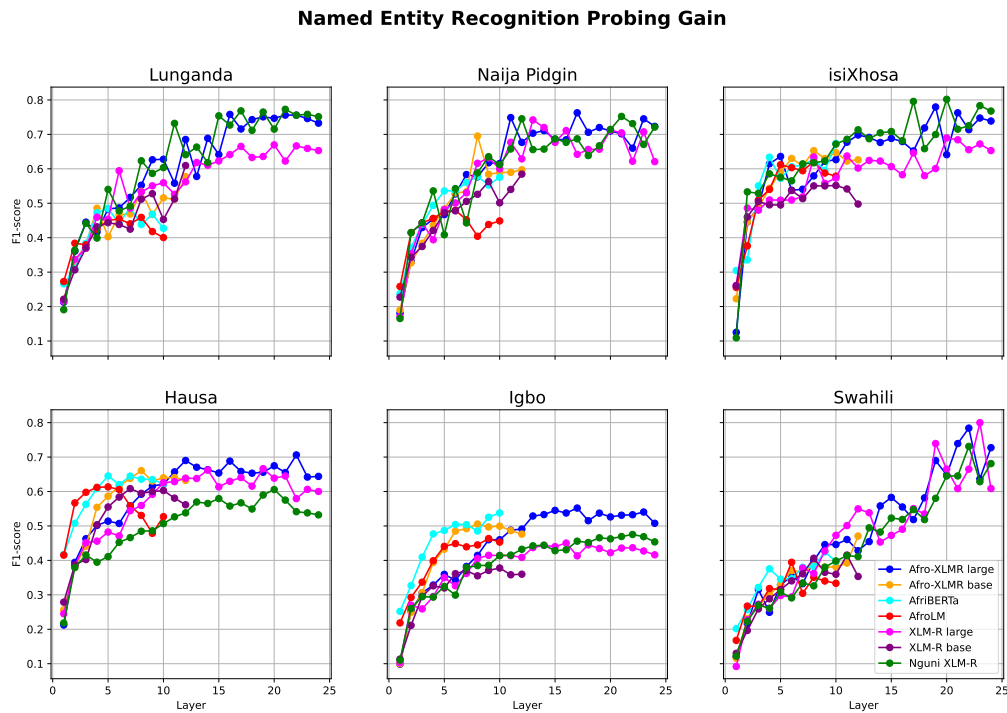


Figure 4.4: Gain for NER across layers

4.4 News Topic Classification

For this task, all models are performing best in isiXhosa as the gap between the performances and the baseline is far wider than in any other language, as shown in Figure 4.5. The figure shows that all models encode topic understanding in all layers, except for Luganda where the performance is noisy across layers. On the other end, all 7 models show their weakest performance in Swahili.

The performance in Naija Pidgin shows the least variation, while Luganda exhibits the most variability across models. As before, later layers generally yield better performance than earlier ones. However, unlike the previous two tasks—where performance starts low in the early layers and peaks towards the middle or later layers—accuracies in this task remain relatively uniform across layers. This suggests that all layers encode a similar amount of knowledge relevant to topic classification, with the exception of Luganda.

Nguni-XLMR, AfriBERTa, and Afro XLMR-large consistently performed well across all languages. As shown in Table 4.3, Afro XLMR-large achieved the highest raw accuracy in Swahili (84.24%) and Hausa (88.33%), AfriBERTa in Igbo (84.02%), and Nguni-XLMR in Luganda (77.27%), isiXhosa (94.61%), and Naija Pidgin (92.76%).

As noted in the methods section, AfroLM’s maximum input length is limited to 256 tokens—half of what the other models support—which may partially explain its lower performance, particularly in Swahili, isiXhosa, and Igbo.

		swa	ibo	hau	lug	xho	pcm
Afro XLMR Large	Best layer	17	16	11	18	14	24
	Accuracy	84.24	82.47	88.33	74.55	92.93	90.79
	Coverage	★★	★★	★★	☆☆	★★	★★
Afro XLMR base	Best layer	10	10	9	11	9	11
	Accuracy	80.46	76.80	86.12	70.00	87.54	90.79
	Coverage	★★	★★	★★	☆☆	★★	★★
AfriBERTa	Best layer	9	10	10	10	4	4
	Accuracy	82.56	84.02	87.38	74.55	81.48	92.76
	Coverage	★★	★★	★★	☆☆	☆☆	★★
AfroLM	Best layer	7	9	4	10	6	9
	Accuracy	76.05	76.29	85.80	69.09	80.47	92.76
	Coverage	★★	★★	★★	★★	★★	★★
XLM-R large	Best layer	15	3, 24	23	1	19	2, 6
	Accuracy	81.93	73.20	87.38	61.82	86.87	92.76
	Coverage	★★	☆☆	★★	☆☆	★★	☆☆
XLM-R base	Best layer	12	12	12	6	12	12
	Accuracy	84.03	74.23	87.38	64.55	82.15	92.11
	Coverage	★★	☆☆	★★	☆☆	★★	☆☆
Nguni-XLMR	Best layer	11	11	7	16	9	24
	Accuracy	80.25	82.47	83.28	77.27	94.61	92.76
	Coverage	★☆☆	★☆☆	★☆☆	☆☆	★★	☆☆

Table 4.3: Accuracy for NTC across 6 languages. Highest accuracy for each language is boldened. 'Coverage' rates the extent of coverage for the language. 0 star: the language was not seen neither in base model or the adapted model. 1 star: The language was included in the base model but not in the adapted model. 2 stars: The model was either pretrained or adapted for the language. Full table is in Appendix A.

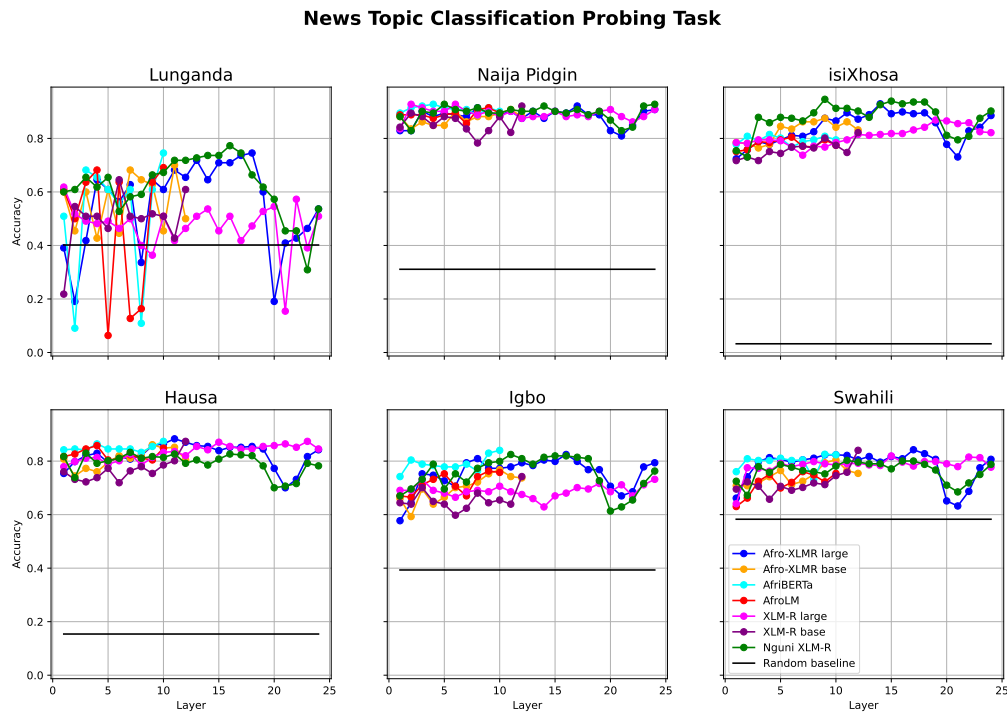


Figure 4.5: News Topic Classification task across all layers

5. Conclusion

5.1 Conclusion

This study reveals the strengths and limitations of current multilingual models in encoding syntactic and semantic knowledge in African languages. While models perform well in languages like Swahili, Hausa, and isiXhosa, the consistent challenges observed with Igbo reinforces the need for more language-specific datasets and training techniques. Additionally, the models' ability to generalize knowledge to low-resource or unseen languages (such as Luganda) points towards a promising direction for further research in cross-lingual transfer learning. Adapting models for specific linguistic groups, as demonstrated with Nguni XLM-R, offers another practical strategy for improving performance.

5.2 Limitations and Future Work

This study is the first of its kind that analyzes language models on their understanding of syntax and semantics of African languages — languages that are severely under-resourced. However, it is important to outline some limitations of the current work in order to provide proper context to the results and findings. These limitations should highlight areas for improvement, and to also point to promising directions for further exploration.

5.2.1 Control Tasks for NER and NTC Datasets

As discussed in Section 3.3.3, designing control tasks for NER proved challenging. While control tasks are primarily designed for word-level tasks, NER presents complications because named entities often span multiple words. This makes it difficult to apply the typical control task framework in a meaningful way. Instead, I relied on random baselines, which, though commonly used, are considered insufficient for robust evaluation (Belinkov, 2022; Hewitt and Liang, 2019). A stronger baseline would have provided stronger evidence for the validity of my NER probing tasks. Developing appropriate control tasks for NER remains an open problem and a promising area for future work. Additionally, extending this effort to news topic classification (NTC) tasks could offer further insights into the limits of current probing methodologies for semantic knowledge.

5.2.2 Unexplored Subword Pooling Strategies

Time constraints limited the exploration of subword pooling strategies, even though the way subwords are aggregated can make a difference in probing performance (Ács et al., 2021). In this study, I used the first subword as input for the classifier to align tokens with their hidden representations. However, other pooling strategies — such as using the last subword, mean pooling, or attention over subwords, could provide different insights, especially for morphologically rich languages with high subword tokenization rates. Future work should systematically compare the effects of different subword pooling strategies across various syntactic and semantic tasks for African languages.

5.2.3 Broader Scope and Additional Interpretability Methods

This work provides valuable insights into the interpretability of language models, but further exploration is needed to enhance our understanding. In future research, I aim to expand beyond the current probing framework by incorporating additional interpretability methods, such as causal probing (Amini

et al., 2023; Ferrando et al., 2024). Causal probing can establish causal relationships between specific contextualized hidden representations and linguistic properties of interest, providing deeper insights into what the model truly captures. This could be especially useful for understanding syntactic and semantic knowledge, as well as the model’s handling of linguistic phenomena unique to a language. The insights would also be useful for targeting specific layers for efficient multilingual-adaptive finetuning.

Acknowledgements

My master's journey, which began in January, has been both demanding and incredibly rewarding, made even more meaningful by the sense of family I found among my coursemates, as well as the academic and non-academic staff.

I am profoundly grateful to Google DeepMind for fully funding my studies at AIMS and for providing me with a mentor, Dr. Luke Maris, whose guidance has been invaluable.

A special thank you goes to my supervisor, François Meyer, for his unwavering support and involvement throughout my project. His technical support and insightful advice have been instrumental in shaping my work from start to finish.

Lastly, I would like to acknowledge Professor Ulrich Paquet and Professor Claire David for their dedication in designing the academic calendar and for bringing in distinguished experts to share their knowledge with us.

References

- Ács, J., Kádár, A., and Kornai, A. Subword pooling makes a difference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2021*. Association for Computational Linguistics, 2021.
- Adelani, D., Neubig, G., Ruder, S., Rijhwani, S., Beukman, M., Palen-Michel, C., Lignos, C., Alabi, J., Muhammad, S., Nabende, P., Dione, C. M. B., Bukula, A., Mabuya, R., Dossou, B. F. P., Sibanda, B., Buzaaba, H., Mukiibi, J., Kalipe, G., Mbaye, D., Taylor, A., Kabore, F., Emezue, C. C., Aremu, A., Ogayo, P., Gitau, C., Munkoh-Buabeng, E., Memdjokam Koagne, V., Tapo, A. A., Macucwa, T., Marivate, V., Elvis, M. T., Gwadabe, T., Adewumi, T., Ahia, O., Nakatumba-Nabende, J., Mokono, N. L., Ezeani, I., Chukwuneke, C., Oluwaseun Adeyemi, M., Hacheme, G. Q., Abdulmumin, I., Ogundepo, O., Yousuf, O., Moteu, T., and Klakow, D. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. URL <https://aclanthology.org/2022.emnlp-main.298>.
- Adelani, D. I., Masiak, M., Azime, I. A., Alabi, J. O., Tonja, A. L., Mwase, C., Ogundepo, O., Dossou, B. F. P., Oladipo, A., Nixdorf, D., Emezue, C. C., Al-Azzawi, S., Sibanda, B. K., David, D., Ndolela, L., Mukiibi, J., Ajayi, T. O., Ngoli, T. M., Odhiambo, B., Owodunni, A. T., Obiefuna, N. C., Muhammad, S. H., Abdullahi, S. S., Yigezu, M. G., Gwadabe, T. R., Abdulmumin, I., Bame, M. T., Awoyomi, O. O., Shode, I., Adelani, T. A., Kailani, H. A., Omotayo, A.-H., Adeeko, A., Abeeb, A., Aremu, A., Samuel, O., Siro, C., Kimotho, W., Ogbu, O. R., Mbonu, C. E., Chukwuneke, C. I., Fanijo, S., Ojo, J., Awosan, O. F., Guge, T. K., Sari, S. T., Nyatsine, P., Sidume, F., Yousuf, O., Oduwole, M., Kimanuka, U. A., Tshinu, K. P., Diko, T., Nxakama, S., Johar, A. T., Gebre, S., Mohamed, M., Mohamed, S. A., Hassan, F. M., Mehamed, M. A., Ngabire, E., and Stenetorp, P. Masakhanews: News topic classification for african languages. 2023.
- Alabi, J. O., Adelani, D. I., Mosbach, M., and Klakow, D. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea, Oct. 2022. International Committee on Computational Linguistics. URL <https://aclanthology.org/2022.coling-1.382>.
- Alain, G. and Bengio, Y. Understanding intermediate layers using linear classifier probes, 2018. URL <https://arxiv.org/abs/1610.01644>.
- Amini, A., Pimentel, T., Meister, C., and Cotterell, R. Naturalistic Causal Probing for Morpho-Syntax. *Transactions of the Association for Computational Linguistics*, 11:384–403, 05 2023. ISSN 2307-387X. doi: 10.1162/tacl_a_00554. URL https://doi.org/10.1162/tacl_a_00554.
- Arora, A., Kaffee, L.-A., and Augenstein, I. Probing pre-trained language models for cross-cultural differences in values. *ArXiv*, abs/2203.13722, 2022. URL <https://api.semanticscholar.org/CorpusID:247748753>.
- Arps, D., Kallmeyer, L., Samih, Y., and Sajjad, H. Multilingual nonce dependency treebanks: Understanding how language models represent and process syntactic structure. In Duh, K., Gomez, H., and Bethard, S., editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*,

- pages 7822–7844, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.433. URL <https://aclanthology.org/2024.naacl-long.433>.
- Belinkov, Y. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1):207–219, 04 2022. ISSN 0891-2017. doi: 10.1162/coli_a_00422. URL https://doi.org/10.1162/coli_a_00422.
- Belinkov, Y., Durrani, N., Dalvi, F., Sajjad, H., and Glass, J. What do neural machine translation models learn about morphology? *arXiv preprint arXiv:1704.03471*, 2017.
- Chrupała, G., Higy, B., and Alishahi, A. Analyzing analytical methods: The case of phonology in neural models of spoken language. *arXiv preprint arXiv:2004.07070*, 2020.
- Conneau, A., Kruszewski, G., Lample, G., Barrault, L., and Baroni, M. What you can cram into a single \$&!#* vector: Probing sentence embeddings for linguistic properties. In Gurevych, I. and Miyao, Y., editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia, July 2018a. Association for Computational Linguistics. doi: 10.18653/v1/P18-1198. URL <https://aclanthology.org/P18-1198>.
- Conneau, A., Kruszewski, G., Lample, G., Barrault, L., and Baroni, M. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*, 2018b.
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. Unsupervised cross-lingual representation learning at scale. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL <https://aclanthology.org/2020.acl-main.747>.
- Dione, C. M. B., Adelani, D. I., Nabende, P., Alabi, J., Sindane, T., Buzaaba, H., Muhammad, S. H., Emezue, C. C., Ogayo, P., Aremu, A., Gitau, C., Mbaye, D., Mukiibi, J., Sibanda, B., Dossou, B. F. P., Bukula, A., Mabuya, R., Tapo, A. A., Munkoh-Buabeng, E., Memdjokam Koagne, V., Ouoba Kabore, F., Taylor, A., Kalipe, G., Macucwa, T., Marivate, V., Gwadabe, T., Elvis, M. T., 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., Uchekukwu, C., Yusuf, A., Abdullahi, M., and Klakow, D. MasakhaPOS: Part-of-speech tagging for typologically diverse African languages. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10883–10900, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.609. URL <https://aclanthology.org/2023.acl-long.609>.
- Dossou, B. F. P., Tonja, A. L., Yousuf, O., Osei, S., Oppong, A., Shode, I., Awoyomi, O. O., and Emezue, C. AfroLM: A self-active learning-based multilingual pretrained language model for 23 African languages. In *Proceedings of The Third Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pages 52–64, Abu Dhabi, United Arab Emirates (Hybrid), Dec. 2022. Association for Computational Linguistics. URL <https://aclanthology.org/2022.sustainlp-1.11>.
- Eberhard, D. M. and Gary, F. Simons, and charles d. fennig (eds.). 2019. *Ethnologue: Languages of the world*, 22.

- Ferrando, J., Sarti, G., Bisazza, A., and Costa-jussà, M. R. A primer on the inner workings of transformer-based language models, 2024. URL <https://arxiv.org/abs/2405.00208>.
- Goyal, N., Du, J., Ott, M., Anantharaman, G., and Conneau, A. Larger-scale transformers for multilingual masked language modeling. *arXiv preprint arXiv:2105.00572*, 2021.
- Hadi, M. U., Al Tashi, Q., Shah, A., Qureshi, R., Muneer, A., Irfan, M., Zafar, A., Shaikh, M. B., Akhtar, N., Wu, J., et al. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*, 2024.
- Hedderich, M. A., Adelani, D., Zhu, D., Alabi, J., Markus, U., and Klakow, D. Transfer learning and distant supervision for multilingual transformer models: A study on African languages. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2580–2591, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.204. URL <https://aclanthology.org/2020.emnlp-main.204>.
- Hewitt, J. and Liang, P. Designing and interpreting probes with control tasks. In *Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2019.
- Hewitt, J. and Manning, C. D. A structural probe for finding syntax in word representations. In Burstein, J., Doran, C., and Solorio, T., editors, *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)*, pages 4129–4138, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1419. URL <https://aclanthology.org/N19-1419>.
- Hou, J., Katinskaia, A., Kotilainen, L., Trangcasanchai, S., Vu, A.-D., and Yangarber, R. What do transformers know about government? In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17459–17472, Torino, Italia, May 2024. ELRA and ICCL. URL <https://aclanthology.org/2024.lrec-main.1518>.
- Katinskaia, A. and Yangarber, R. Probing the category of verbal aspect in transformer language models. In Duh, K., Gomez, H., and Bethard, S., editors, *Findings of the Association for Computational Linguistics*, pages 3347–3366, United States, 2024. The Association for Computational Linguistics. Findings of the Association for Computational Linguistics : NAACL 2024 ; Conference date: 16-06-2024 Through 21-06-2024.
- Köhn, A. What’s in an embedding? analyzing word embeddings through multilingual evaluation. 2015.
- Li, D., Jin, M., Zeng, Q., Zhao, H., and Du, M. Exploring multilingual probing in large language models: A cross-language analysis, 2024. URL <https://arxiv.org/abs/2409.14459>.
- Liu, N. F., Gardner, M., Belinkov, Y., Peters, M. E., and Smith, N. A. Linguistic knowledge and transferability of contextual representations. *arXiv preprint arXiv:1903.08855*, 2019.
- Lovering, C., Jha, R., Linzen, T., and Pavlick, E. Predicting inductive biases of pre-trained models. In *ICLR*. OpenReview.net, 2021. URL <http://dblp.uni-trier.de/db/conf/iclr/iclr2021.html#LoveringJLP21>.
- Luo, H. and Specia, L. From understanding to utilization: A survey on explainability for large language models, 2024. URL <https://arxiv.org/abs/2401.12874>.

- Meyer, F., Song, H., Chakrabarty, A., Buys, J., Dabre, R., and Tanaka, H. NGLUEni: Benchmarking and adapting pretrained language models for nguni languages. In Calzolari, N., Kan, M.-Y., Hoste, V., Lenci, A., Sakti, S., and Xue, N., editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12247–12258, Torino, Italia, May 2024. ELRA and ICCL. URL <https://aclanthology.org/2024.lrec-main.1071>.
- Michael, J., Botha, J. A., and Tenney, I. Asking without telling: Exploring latent ontologies in contextual representations. *arXiv preprint arXiv:2004.14513*, 2020.
- Millière, R. Language models as models of language, 2024. URL <https://arxiv.org/abs/2408.07144>.
- Ogueji, K., Zhu, Y., and Lin, J. 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*, pages 116–126, Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics. URL <https://aclanthology.org/2021.mrl-1.11>.
- Pimentel, T., Valvoda, J., Maudslay, R. H., Zmigrod, R., Williams, A., and Cotterell, R. Information-theoretic probing for linguistic structure. *arXiv preprint arXiv:2004.03061*, 2020.
- Ravichander, A., Belinkov, Y., and Hovy, E. Probing the probing paradigm: Does probing accuracy entail task relevance? *arXiv preprint arXiv:2005.00719*, 2020.
- Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Van Durme, B., Bowman, S. R., Das, D., et al. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*, 2019.
- Vanmassenhove, E., Du, J., and Way, A. Investigating ‘aspect’ in nmt and smt: translating the english simple past and present perfect. *Computational Linguistics in the Netherlands Journal (CLIN)*, 7: 109–128, 2017.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.
- Zhang, C., Tong, H., Zhang, B., and Zhang, D. Probing causality manipulation of large language models, 2024. URL <https://arxiv.org/abs/2408.14380>.
- Zhang, K. and Bowman, S. Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis. In Linzen, T., Chrupała, G., and Alishahi, A., editors, *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 359–361, Brussels, Belgium, Nov. 2018a. Association for Computational Linguistics. doi: 10.18653/v1/W18-5448. URL <https://aclanthology.org/W18-5448>.
- Zhang, K. W. and Bowman, S. R. Language modeling teaches you more syntax than translation does: Lessons learned through auxiliary task analysis. *arXiv preprint arXiv:1809.10040*, 2018b.

.1 Appendix A

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Afro-XLMR large	84.24	88.81	89.96	90.74	90.72	90.97	91.75	91.86	91.99	92.20	92.25	91.98	91.41	91.05	90.76	90.91	90.72	90.49	90.30	89.98	90.10	89.26	89.00	89.62
Afro-XLMR base	84.45	89.75	91.16	92.13	92.47	92.16	92.17	92.08	92.16	91.97	91.71	91.81	-	-	-	-	-	-	-	-	-	-	-	-
AfriBERTa	87.70	89.73	92.29	92.67	92.39	92.39	92.14	92.12	91.88	91.37	-	-	-	-	-	-	-	-	-	-	-	-	-	-
AfroLM	83.57	86.34	90.47	91.70	91.64	91.38	91.02	90.81	89.94	89.72	-	-	-	-	-	-	-	-	-	-	-	-	-	-
XLM-R large	84.50	89.19	89.49	89.99	89.86	90.00	90.84	91.17	91.34	91.66	91.68	91.43	90.90	91.16	90.48	90.09	89.84	89.73	89.88	89.75	89.78	89.89	89.75	89.38
XLM-R base	84.43	88.43	90.76	91.33	91.76	91.82	91.82	91.58	91.68	91.58	91.44	90.72	-	-	-	-	-	-	-	-	-	-	-	-
Nguni XLM-R	84.27	87.64	88.88	88.82	89.37	89.54	89.45	90.29	90.56	90.41	90.56	90.04	90.25	90.22	89.75	89.88	89.59	89.60	89.62	89.85	89.32	89.25	89.47	89.32

Table 1: POS results for Swahili. Accuracy reported.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Afro-XLMR large	69.78	82.48	84.11	84.61	84.69	84.52	84.40	84.55	86.01	86.24	86.28	86.60	85.98	85.93	85.79	85.78	85.19	84.67	85.08	85.53	85.13	85.47	84.45	84.69
Afro-XLMR base	69.89	81.05	84.67	86.16	86.72	87.11	87.42	87.42	87.35	87.66	87.87	86.56	-	-	-	-	-	-	-	-	-	-	-	-
AfriBERTa	68.63	72.38	82.22	85.44	85.58	85.90	85.25	85.39	84.87	83.89	-	-	-	-	-	-	-	-	-	-	-	-	-	-
AfroLM	72.85	75.33	82.47	86.84	87.68	87.03	86.11	85.28	84.24	83.39	-	-	-	-	-	-	-	-	-	-	-	-	-	-
XLM-R large	70.15	81.56	83.00	83.39	83.35	82.39	82.03	80.77	80.66	80.06	79.01	78.43	78.49	76.75	76.44	76.27	75.87	75.71	75.55	76.51	77.73	77.65	77.58	77.03
XLM-R base	69.69	79.18	81.05	81.87	81.42	81.64	80.99	81.39	80.85	82.12	81.44	80.23	-	-	-	-	-	-	-	-	-	-	-	-
Nguni XLM-R	69.99	81.82	83.68	84.19	84.57	84.11	84.61	85.79	86.85	87.14	87.11	86.66	86.26	86.78	86.04	85.59	85.28	85.65	86.05	86.25	86.66	86.58	86.43	86.65

Table 2: POS results for Luganda. Accuracy reported.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Afro-XLMR large	81.70	88.54	89.90	90.16	90.09	90.46	90.99	91.28	91.78	91.74	91.80	91.78	91.43	90.98	91.31	90.80	90.90	90.42	90.70	90.74	90.77	90.72	90.41	90.18
Afro-XLMR base	80.93	89.07	90.62	91.83	91.91	92.34	92.17	92.34	92.55	92.24	92.31	92.17	-	-	-	-	-	-	-	-	-	-	-	-
AfriBERTa	87.55	89.41	92.28	92.67	92.50	92.13	92.06	92.10	91.80	91.32	-	-	-	-	-	-	-	-	-	-	-	-	-	-
AfroLM	89.70	90.76	91.51	91.80	91.08	89.99	87.90	87.01	85.77	84.91	-	-	-	-	-	-	-	-	-	-	-	-	-	-
XLM-R large	81.32	88.18	88.51	88.98	88.93	89.00	89.50	90.07	90.71	90.99	91.07	90.59	90.35	90.07	89.74	89.94	89.56	89.47	89.71	89.80	89.65	89.41	89.50	89.43
XLM-R base	81.04	87.26	89.43	90.13	90.26	90.63	90.69	91.03	90.47	90.72	90.14	90.14	-	-	-	-	-	-	-	-	-	-	-	-
Nguni XLM-R	81.38	87.48	88.07	88.03	88.29	88.31	88.51	88.22	88.18	87.68	87.46	87.28	86.83	86.76	86.44	85.72	85.39	85.55	85.38	85.96	86.73	86.27	86.43	85.97

Table 3: POS results for Hausa. Accuracy reported.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Afro-XLMR large	83.03	86.06	87.10	87.27	87.25	87.20	87.61	88.26	88.90	89.26	89.16	89.14	88.96	88.90	88.72	88.60	88.75	88.49	88.58	88.57	88.52	88.55	88.28	87.82
Afro-XLMR base	82.60	86.40	87.50	88.83	89.62	89.61	89.69	89.77	89.68	89.66	89.53	89.28	-	-	-	-	-	-	-	-	-	-	-	-
AfriBERTa	82.29	84.65	87.24	88.19	88.34	88.14	88.40	88.09	87.65	87.39	-	-	-	-	-	-	-	-	-	-	-	-	-	-
AfroLM	84.90	86.56	87.38	87.72	87.57	86.74	85.51	84.48	83.12	82.13	-	-	-	-	-	-	-	-	-	-	-	-	-	-
XLM-R large	83.22	86.27	86.78	87.02	86.85	86.94	87.19	87.45	87.62	87.59	87.63	87.55	87.11	86.69	86.58	86.08	85.16	85.19	85.70	85.51	85.43	85.12	85.11	85.05
XLM-R base	83.10	86.02	86.66	87.54	87.52	87.38	87.33	87.17	87.40	87.12	87.17	86.27	-	-	-	-	-	-	-	-	-	-	-	-
Nguni XLM-R	82.65	86.33	86.62	86.56	86.84	86.90	87.15	87.42	87.70	87.70	87.45	87.36	87.27	87.04	87.12	86.96	86.88	86.88	86.78	87.04	86.69	86.67	86.69	86.10

Table 4: POS results for Naija Pidgin. Accuracy reported.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Afro-XLMR large	70.69	79.74	82.32	82.71	83.54	83.05	84.09	85.19	85.70	86.19	86.10	85.49	84.74	84.74	84.39	84.29	83.91	83.24	83.96	83.75	84.21	84.02	83.40	83.59
Afro-XLMR base	70.33	80.17	84.15	86.31	86.79	86.74	86.87	87.32	86.90	87.08	86.88	86.59	-	-	-	-	-	-	-	-	-	-	-	-
AfriBERTa	64.15	64.69	78.56	81.41	81.76	81.39	81.41	81.15	80.48	80.06	-	-	-	-	-	-	-	-	-	-	-	-	-	-
AfroLM	73.90	74.33	80.53	84.42	84.74	84.78	83.93	82.96	82.91	82.65	-	-	-	-	-	-	-	-	-	-	-	-	-	-
XLM-R large	69.66	78.86	80.04	80.53	81.01	80.31	81.37	82.45	82.95	82.79	82.68	81.81	81.79	81.02	80.44	79.87	80.83	80.38	79.89	80.40	80.49	80.06	80.14	80.48
XLM-R base	70.17	77.74	80.21	81.68	82.37	82.37	82.13	83.19	82.01	82.70	82.01	82.15	-	-	-	-	-	-	-	-	-	-	-	-
Nguni XLM-R	71.10	81.43	84.07	84.82	84.65	84.73	85.00	85.88	86.29	86.11	85.99	84.82	84.41	84.24	84.00	83.98	82.94	83.09	83.36	83.76	83.86	83.90	83.57	83.01

Table 5: POS results for isiXhosa. Accuracy reported.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Afro-XLMR large	61.64	71.46	73.09	73.44	74.33	74.19	74.71	75.45	76.20	76.27	76.16	76.28	76.70	76.35	76.18	76.33	76.64	76.48	74.76	74.90	75.52	75.55	74.49	75.30
Afro-XLMR base	62.13	72.67	75.08	76.47	77.07	77.07	78.00	77.45	77.51	77.32	77.59	77.76	-	-	-	-	-	-	-	-	-	-	-	-
AfriBERTa	71.00	72.46	76.94	76.95	78.12	77.98	77.97	77.47	78.01	77.74	-	-	-	-	-	-	-	-	-	-	-	-	-	-
AfroLM	66.48	67.34	73.35	75.35	75.04	76.10	75.77	75.53	74.76	75.30	-	-	-	-	-	-	-	-	-	-	-	-	-	-
XLM-R large	61.47	71.21	72.51	71.46	72.84	72.23	72.20	72.94	71.74	71.03	72.02	71.52	70.04	70.28	70.50	71.77	69.87	71.90	71.23	71.46	71.79	70.59	70.25	68.63
XLM-R base	63.90	71.32	73.00	72.85	73.26	73.15	69.58	70.29	71.92	68.96	71.55	71.58	-	-	-	-	-	-	-	-	-	-	-	-
Nguni XLM-R	58.54	71.47	72.37	72.45	72.28	72.01	72.12	72.12	72.40	71.21	72.52	72.14	72.45	71.97	70.77	71.71	71.97	71.24	71.17	70.09	70.45	70.21	70.08	68.76

Table 6: POS results for Igbo. Accuracy reported.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Afro-XLMR large	5.25	7.31	8.16	8.87	9.74	9.15	10.56	12.80	13.59	13.66	15.50	17.57	18.68	19.12	19.63	20.93	21.82	21.38	20.46	19.80	20.18	18.72	19.11	18.88
Afro-XLMR base	5.47	8.94	10.16	12.12	13.63	15.35	16.46	16.73	16.06	15.48	14.78	16.36	-	-	-	-	-	-	-	-	-	-	-	-
AfriBERTa	5.59	8.35	10.54	12.14	13.52	16.29	16.88	17.46	18.68	18.79	-	-	-	-	-	-	-	-	-	-	-	-	-	-
AfroLM	8.47	13.52	14.03	14.23	16.60	20.16	22.94	23.58	23.64	23.27	-	-	-	-	-	-	-	-	-	-	-	-	-	-
XLM-R large	5.22	7.69	7.52	8.19	8.64	8.47	10.48	10.93	11.68	12.87	14.43	15.52	17.16	18.58	18.24	17.69	19.09	18.70	19.09	18.51	17.23	17.38	16.57	15.97
XLM-R base	5.03	8.30	10.34	12.01	12.92	14.89	15.09	15.95	15.51	14.77	15.83	16.39	-	-	-	-	-	-	-	-	-	-	-	-
Nguni XLM-R	6.38	6.07	7.04	7.78	8.67	8.94	9.94	10.66	12.16	12.81	14.25	14.55	17.01	18.15	17.72	19.09	18.14	19.38	18.70	18.73	17.41	18.19	18.36	17.51

Table 7: POS results for Swahili. Selectivity reported.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Afro-XLMR large	12.28	14.28	13.32	13.47	15.11	15.65	16.15	19.01	19.29	20.01	21.44	23.39	23.41	24.63	24.89	25.82	26.38	26.71	26.97	26.71	25.75	24.90	25.08	24.44
Afro-XLMR base	13.01	15.49	16.11	17.52	20.34	21.46	23.01	23.07	22.22	23.47	24.18	22.57	-	-	-	-	-	-	-	-	-	-	-	-
AfriBERTa	14.15	17.41	17.23	17.12	17.87	16.93	18.45	18.56	17.91	18.85	-	-	-	-	-	-	-	-	-	-	-	-	-	-
AfroLM	13.67	16.15	17.75	20.52	24.55	26.88	27.66	27.98	26.94	27.76	-	-	-	-	-	-	-	-	-	-	-	-	-	-
XLM-R large	12.35	14.75	14.51	14.84	14.63	15.68	15.22	16.51	16.94	16.94	17.66	19.05	19.75	19.16	18.68	19.97	19.81	19.77	19.92	19.46	19.72	18.21	18.27	19.14
XLM-R base	12.11	15.89	14.65	16.89	17.75	18.57	19.06	19.23	18.29	17.99	18.82	19.70	-	-	-	-	-	-	-	-	-	-	-	-
Nguni XLM-R	12.70	13.48	12.24	13.74	13.77	15.13	16.85	17.70	18.87	20.93	21.64	23.33	24.14	25.52	24.68	25.67	25.21	25.36	25.41	23.75	24.29	23.72	23.48	24.08

Table 8: POS results for Luganda. Selectivity reported.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Afro-XLMR large	3.41	3.98	4.20	3.85	4.98	4.89	5.67	6.58	8.53	9.50	9.98	12.38	12.63	13.08	14.91	14.84	16.02	15.91	15.92	16.35	14.88	16.56	13.85	13.14
Afro-XLMR base	3.06	5.49	5.60	7.42	9.65	11.17	13.05	12.89	14.40	12.45	13.28	12.58	-	-	-	-	-	-	-	-	-	-	-	-
AfriBERTa	2.16	4.62	7.01	8.16	10.65	11.71	13.05	14.77	14.05	13.57	-	-	-	-	-	-	-	-	-	-	-	-	-	-
AfroLM	1.20	4.14	7.33	12.61	16.72	22.46	24.76	26.55	25.63	26.17	-	-	-	-	-	-	-	-	-	-	-	-	-	-
XLM-R large	2.92	4.79	4.33	4.06	3.60	4.50	5.53	5.88	6.85	9.20	9.89	10.50	12.47	12.39	13.62	14.20	14.46	14.54	14.52	13.72	12.74	11.32	11.02	11.53
XLM-R base	3.09	5.45	7.14	7.10	9.61	11.05	11.99	12.43	12.92	11.60	12.17	13.15	-	-	-	-	-	-	-	-	-	-	-	-
Nguni XLM-R	3.84	2.95	2.38	2.52	3.18	4.07	3.71	4.70	5.95	5.47	6.36	7.63	7.52	9.39	8.55	7.95	9.48	9.60	9.25	9.76	8.37	7.63	7.97	7.48

Table 9: POS results for Hausa. Selectivity reported.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Afro-XLMR large	3.35	1.50	2.43	3.94	5.10	5.16	7.34	8.37	10.92	12.31	13.76	15.40	16.96	18.21	19.62	20.50	21.57	21.89	22.03	22.10	21.75	19.89	20.40	18.98
Afro-XLMR base	2.76	4.00	6.80	8.97	10.70	12.97	16.28	15.41	15.90	15.76	15.27	15.58	-	-	-	-	-	-	-	-	-	-	-	-
AfriBERTa	0.81	3.82	4.36	6.36	9.38	10.77	12.15	13.25	14.03	14.55	-	-	-	-	-	-	-	-	-	-	-	-	-	-
AfroLM	0.86	4.86	5.52	11.78	16.71	21.91	25.98	27.28	26.96	27.08	-	-	-	-	-	-	-	-	-	-	-	-	-	-
XLM-R large	1.99	2.26	2.26	3.36	3.66	4.97	5.83	9.10	9.96	10.96	12.28	13.82	15.08	16.40	16.56	17.92	17.47	19.56	18.42	19.51	17.36	16.72	16.37	16.68
XLM-R base	2.40	3.18	4.89	7.56	8.63	10.67	11.42	12.47	14.39	12.33	15.29	15.39	-	-	-	-	-	-	-	-	-	-	-	-
Nguni XLM-R	2.39	2.24	2.61	3.92	4.80	4.91	7.07	9.11	9.84	10.03	11.80	12.77	14.09	14.87	16.73	17.63	18.19	18.46	18.57	18.92	16.20	16.17	15.68	15.33

Table 10: POS results for Naija Pidgin. Selectivity reported.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Afro-XLMR large	26.05	29.89	29.18	30.31	30.74	32.20	33.13	34.61	34.55	35.61	37.32	37.82	39.11	39.76	40.91	40.85	42.49	39.55	39.92	40.06	40.85	40.08	42.12	37.88
Afro-XLMR base	25.85	30.63	31.99	33.64	35.76	35.38	36.99	38.48	37.19	38.26	37.59	37.64	-	-	-	-	-	-	-	-	-	-	-	-
AfriBERTa	22.70	22.94	30.87	29.71	29.16	29.93	28.78	30.48	27.56	28.49	-	-	-	-	-	-	-	-	-	-	-	-	-	-
AfroLM	28.08	28.45	31.26	34.51	35.93	38.01	38.92	38.62	38.76	38.81	-	-	-	-	-	-	-	-	-	-	-	-	-	-
XLM-R large	24.35	28.89	29.92	30.07	32.15	30.86	31.43	33.39	34.38	36.06	35.82	35.35	36.73	37.21	37.03	37.25	38.15	37.61	37.30	37.81	36.19	36.39	36.19	36.86
XLM-R base	25.85	30.31	31.89	33.09	34.44	34.98	33.85	36.32	35.15	36.59	35.49	36.24	-	-	-	-	-	-	-	-	-	-	-	-
Nguni XLM-R	26.68	30.08	31.83	33.20	31.97	33.76	33.14	34.96	36.89	37.35	38.10	36.75	39.10	43.13	42.25	41.67	41.09	40.27	41.45	40.24	41.90	39.64	42.54	37.83

Table 11: POS results for isiXhosa. Selectivity reported.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Afro-XLMR large	3.43	-2.66	-4.85	-4.75	-4.04	-3.11	-2.74	-1.15	-1.27	-0.13	0.61	2.13	3.37	4.17	5.27	5.81	6.58	7.62	6.63	6.40	6.17	7.08	4.38	4.63
Afro-XLMR base	4.22	-1.47	-3.63	-2.73	-1.55	-1.80	1.29	1.02	1.03	0.05	0.27	0.82	-	-	-	-	-	-	-	-	-	-	-	-
AfriBERTa	-10.02	-8.78	-6.80	-6.72	-4.25	-3.24	-2.39	-2.67	-1.58	-0.82	-	-	-	-	-	-	-	-	-	-	-	-	-	-
AfroLM	-1.37	0.00	-3.16	-3.08	-2.72	0.47	1.57	1.36	1.61	3.37	-	-	-	-	-	-	-	-	-	-	-	-	-	-
XLM-R large	3.29	-2.60	-3.55	-4.22	-4.27	-3.09	-3.69	-1.68	-2.69	-2.18	-0.28	0.64	-1.18	0.44	0.96	2.31	0.80	3.48	2.38	1.72	1.01	-0.57	-0.58	-0.95
XLM-R base	5.98	-1.07	-2.70	-3.58	-1.30	-1.23	-4.83	-2.91	-2.54	-6.64	-3.32	-0.96	-	-	-	-	-	-	-	-	-	-	-	-
Nguni XLM-R	0.73	-3.08	-4.61	-3.91	-4.55	-4.47	-4.23	-2.75	-3.30	-3.23	0.23	0.44	2.02	1.92	0.92	2.13	3.33	3.35	3.60	1.91	1.61	1.22	1.37	0.89

Table 12: POS results for Igbo. Selectivity reported.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Afro-XLMR large	68.27	78.78	81.56	81.94	82.78	83.11	84.20	85.33	86.42	86.75	87.72	87.90	86.31	86.73	86.19	83.78	84.01	83.04	84.16	83.43	83.24	82.08	82.76	81.80
Afro-XLMR base	67.80	78.40	80.85	84.76	86.12	87.94	87.91	88.37	88.22	87.86	87.54	86.15	-	-	-	-	-	-	-	-	-	-	-	-
AfriBERTa	76.07	80.63	85.90	87.57	88.54	88.94	88.86	88.46	88.51	87.98	-	-	-	-	-	-	-	-	-	-	-	-	-	-
AfroLM	73.63	80.71	82.25	84.52	85.66	85.83	85.01	84.77	83.79	83.83	-	-	-	-	-	-	-	-	-	-	-	-	-	-
XLM-R large	67.46	77.72	79.80	80.71	81.23	81.77	83.09	84.36	85.50	85.46	85.82	85.95	84.90	82.87	83.24	81.96	82.93	82.05	82.49	81.90	82.07	81.82	81.36	80.96
XLM-R base	68.86	76.80	80.31	83.01	84.70	86.35	86.44	86.96	86.33	85.73	85.42	83.29	-	-	-	-	-	-	-	-	-	-	-	-
Nguni XLM-R	67.31	77.50	78.92	79.10	81.31	80.88	82.56	82.52	83.66	83.71	84.31	84.43	84.19	83.77	84.11	84.34	84.04	84.19	84.42	84.81	84.18	83.78	82.42	82.04

Table 13: NER results for Swahili. F1-scores are reported.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Afro-XLMR large	51.86	68.78	70.20	70.37	71.88	72.10	74.16	74.50	76.38	75.45	76.36	75.10	75.58	75.18	75.12	75.88	74.92	75.16	75.09	75.46	75.60	75.59	75.47	73.26
Afro-XLMR base	53.17	66.91	69.78	73.26	74.48	75.46	76.08	77.36	77.29	77.25	76.32	76.14	-	-	-	-	-	-	-	-	-	-	-	-
AfriBERTa	57.92	64.00	70.52	73.69	75.17	74.94	75.32	74.45	75.05	73.54	-	-	-	-	-	-	-	-	-	-	-	-	-	-
AfroLM	63.06	68.75	72.41	75.45	74.82	73.89	71.07	70.47	68.13	67.69	-	-	-	-	-	-	-	-	-	-	-	-	-	-
XLM-R large	52.51	67.17	68.51	70.32	70.74	70.04	71.40	71.10	72.80	71.70	71.74	72.18	71.09	69.49	69.20	67.74	66.50	66.97	67.36	66.95	66.28	66.67	65.92	65.30
XLM-R base	54.02	65.03	68.46	72.16	72.03	72.59	73.25	72.74	72.79	70.89	70.53	68.07	-	-	-	-	-	-	-	-	-	-	-	-
Nguni XLM-R	51.10	69.17	70.72	70.67	70.97	71.33	72.27	73.34	74.58	75.44	75.42	75.03	74.97	76.05	75.38	76.31	76.83	76.68	76.49	76.22	77.28	75.56	75.83	75.14

Table 14: NER results for Luganda. F1-scores are reported.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Afro-XLMR large	36.46	51.58	55.71	60.73	60.61	60.66	65.39	66.50	69.76	71.04	73.27	75.15	74.27	72.75	71.45	72.46	71.65	70.32	72.45	71.87	71.80	71.57	69.61	65.74
Afro-XLMR base	37.80	49.89	56.21	65.59	69.21	72.55	72.47	73.79	73.44	75.01	73.26	70.61	-	-	-	-	-	-	-	-	-	-	-	-
AfriBERTa	55.55	63.11	71.38	73.71	75.76	73.80	75.06	73.52	73.13	71.23	-	-	-	-	-	-	-	-	-	-	-	-	-	-
AfroLM	54.04	67.94	71.45	73.05	73.90	71.26	66.72	63.06	60.33	61.82	-	-	-	-	-	-	-	-	-	-	-	-	-	-
XLM-R large	39.61	52.08	54.77	56.76	57.95	57.78	62.21	63.68	67.34	68.78	69.47	70.36	70.16	69.12	68.37	69.33	68.20	68.70	69.09	69.83	66.85	64.70	62.60	61.49
XLM-R base	39.33	50.37	51.46	62.35	65.48	69.46	70.65	69.93	70.74	68.40	67.96	65.58	-	-	-	-	-	-	-	-	-	-	-	-
Nguni XLM-R	35.38	50.28	51.89	50.61	50.65	54.05	54.14	54.64	57.10	58.55	60.67	63.02	61.55	63.09	63.29	62.23	62.74	61.58	61.52	62.65	61.42	59.74	59.46	55.89

Table 15: NER results for Hausa. F1-scores are reported.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Afro-XLMR large	49.09	64.35	67.03	67.16	68.61	68.07	71.97	71.12	73.99	74.23	76.47	76.71	77.61	77.83	77.58	77.70	76.27	76.80	76.81	76.58	76.36	74.73	74.51	72.33
Afro-XLMR base	46.11	62.25	66.29	73.14	73.70	78.19	79.44	78.88	79.73	80.09	79.39	76.36	-	-	-	-	-	-	-	-	-	-	-	-
AfriBERTa	57.14	66.42	72.19	77.06	78.60	79.30	77.92	77.73	77.87	78.21	-	-	-	-	-	-	-	-	-	-	-	-	-	-
AfroLM	60.60	70.13	73.25	75.24	76.40	73.54	70.92	68.39	67.71	65.76	-	-	-	-	-	-	-	-	-	-	-	-	-	-
XLM-R large	48.42	64.15	68.80	67.52	68.31	68.28	69.95	71.38	72.93	74.80	74.81	75.56	74.46	72.01	73.40	72.54	72.74	72.53	72.03	71.23	70.44	69.17	70.72	68.21
XLM-R base	49.09	63.25	66.00	70.72	73.16	73.39	75.41	75.98	75.37	74.35	74.92	73.30	-	-	-	-	-	-	-	-	-	-	-	-
Nguni XLM-R	48.10	63.94	66.69	67.94	67.52	67.53	70.46	70.71	72.66	72.23	73.09	75.54	75.53	76.78	76.55	74.50	75.25	74.83	74.63	75.79	75.22	73.13	72.44	72.13

Table 16: NER results for Naija Pidgin. F1-scores reported.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Afro-XLMR large	29.36	61.96	65.94	67.87	68.56	67.52	69.39	72.44	74.45	74.87	76.37	77.06	79.01	78.27	78.00	78.67	77.23	77.85	77.96	77.74	76.26	73.64	74.75	73.88
Afro-XLMR base	36.14	61.23	66.27	69.91	73.83	77.33	77.65	78.54	79.46	78.18	77.16	73.61	-	-	-	-	-	-	-	-	-	-	-	-
AfriBERTa	33.23	36.77	63.57	68.08	69.20	69.55	68.57	69.84	68.11	68.47	-	-	-	-	-	-	-	-	-	-	-	-	-	-
AfroLM	50.69	54.67	66.18	72.34	74.63	74.36	74.22	74.00	70.84	70.83	-	-	-	-	-	-	-	-	-	-	-	-	-	-
XLM-R large	33.41	60.38	63.66	65.02	65.07	65.03	67.70	67.20	69.19	70.38	71.09	70.87	71.12	71.66	70.46	69.12	68.94	69.05	70.17	68.97	68.92	68.36	67.21	65.29
XLM-R base	32.99	60.09	62.66	64.92	66.09	67.67	67.38	68.08	67.64	68.00	68.49	64.66	-	-	-	-	-	-	-	-	-	-	-	-
Nguni XLM-R	31.33	65.87	68.81	69.14	70.66	71.03	73.86	74.99	79.04	78.24	80.07	80.89	81.41	81.12	82.23	79.50	79.92	77.44	79.27	80.18	80.20	80.73	78.34	76.78

Table 17: NER results for isiXhosa. F1-scores reported.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Afro-XLMR large	44.07	59.28	63.83	65.08	66.24	65.11	68.71	69.92	72.35	75.46	76.24	78.35	79.87	78.77	79.81	78.78	77.89	76.87	76.88	77.74	76.47	76.01	76.36	73.16
Afro-XLMR base	44.30	56.98	65.11	73.22	76.36	78.94	79.93	79.67	80.86	79.98	79.76	76.63	-	-	-	-	-	-	-	-	-	-	-	-
AfriBERTa	58.71	66.73	75.77	79.55	82.48	82.58	82.05	81.65	81.41	81.82	-	-	-	-	-	-	-	-	-	-	-	-	-	-
AfroLM	55.40	63.18	67.28	74.70	76.72	76.05	76.01	75.71	75.21	74.65	-	-	-	-	-	-	-	-	-	-	-	-	-	-
XLM-R large	45.15	58.80	60.63	61.82	64.81	63.30	66.73	67.59	68.98	68.75	69.98	69.11	69.56	69.21	69.44	69.00	68.32	67.97	68.88	67.69	67.36	66.41	65.64	63.62
XLM-R base	46.29	55.92	61.57	65.83	64.46	68.01	68.68	68.26	68.48	67.49	66.77	62.95	-	-	-	-	-	-	-	-	-	-	-	-
Nguni XLM-R	45.04	58.98	61.30	61.46	63.84	62.34	66.82	66.97	67.27	69.35	68.54	68.90	69.87	69.30	69.89	69.98	69.26	69.57	68.85	69.54	69.84	69.91	68.70	67.44

Table 18: NER results for Igbo. F1-scores reported.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Afro-XLMR large	12.03	21.57	31.55	24.93	32.88	35.40	33.24	39.93	44.59	44.61	46.07	42.82	45.45	55.83	58.29	55.51	51.85	58.16	68.99	64.80	73.92	78.41	63.70	72.73
Afro-XLMR base	11.57	23.08	26.82	30.06	33.04	37.01	36.62	38.43	38.03	38.11	39.22	47.07	-	-	-	-	-	-	-	-	-	-	-	-
AfriBERTa	20.17	25.69	32.16	37.55	34.55	34.67	38.01	38.24	42.45	39.70	-	-	-	-	-	-	-	-	-	-	-	-	-	-
AfroLM	16.77	26.74	26.63	31.83	31.55	39.45	30.43	35.08	34.01	33.35	-	-	-	-	-	-	-	-	-	-	-	-	-	-
XLM-R large	9.19	22.74	27.29	26.01	29.79	29.40	37.77	36.20	42.74	47.29	50.11	54.97	53.83	45.33	47.25	49.04	55.04	53.48	73.93	66.51	60.84	66.49	79.98	60.87
XLM-R base	13.01	19.65	25.90	28.85	31.51	34.05	36.09	40.63	36.54	35.98	41.54	35.34	-	-	-	-	-	-	-	-	-	-	-	-
Nguni XLM-R	12.19	22.30	27.02	26.10	30.80	29.10	33.46	32.60	38.03	39.83	41.37	41.11	49.46	48.22	52.29	51.90	54.76	51.83	58.03	64.51	64.53	73.10	62.85	68.11

Table 19: NER results for Swahili with baselines. F1-gains reported.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Afro-XLMR large	21.29	36.10	44.51	41.63	48.36	48.66	51.71	55.24	62.64	62.77	55.82	68.52	57.78	68.86	64.15	75.79	71.56	74.28	75.09	74.68	75.60	75.59	74.59	73.26
Afro-XLMR base	22.09	33.06	38.03	48.55	40.28	46.68	46.88	53.46	46.75	51.58	51.15	57.70	-	-	-	-	-	-	-	-	-	-	-	-
AfriBERTa	26.59	31.04	38.35	47.27	48.48	44.15	48.75	43.84	46.82	42.73	-	-	-	-	-	-	-	-	-	-	-	-	-	-
AfroLM	27.28	38.39	37.93	42.48	45.00	45.45	44.04	45.90	41.78	40.05	-	-	-	-	-	-	-	-	-	-	-	-	-	-
XLM-R large	21.54	33.73	36.94	45.96	45.34	59.44	48.57	53.26	55.06	55.94	52.63	56.20	61.82	61.05	62.29	64.13	66.50	63.28	63.58	66.95	62.25	66.67	65.92	65.30
XLM-R base	22.12	30.66	37.03	43.15	44.36	43.83	42.48	51.20	52.77	45.30	51.26	60.95	-	-	-	-	-	-	-	-	-	-	-	-
Nguni XLM-R	19.09	36.36	44.21	39.90	54.03	47.77	49.15	62.33	58.66	60.35	73.20	64.11	66.31	61.80	75.38	72.70	76.83	71.17	76.49	71.53	77.28	75.56	75.83	75.14

Table 20: NER results for Luganda with baselines. F1-gains reported.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Afro-XLMR large	21.25	39.43	46.29	50.17	51.40	50.72	55.84	59.22	61.57	62.15	65.71	69.01	67.05	66.33	65.38	68.84	65.83	65.33	65.60	67.47	65.51	70.63	64.20	64.39
Afro-XLMR base	25.64	38.12	44.11	55.41	58.64	61.55	63.81	66.06	63.05	64.00	64.05	63.22	-	-	-	-	-	-	-	-	-	-	-	-
AfriBERTa	41.72	50.77	56.30	60.87	64.52	62.03	64.46	63.58	63.49	62.27	-	-	-	-	-	-	-	-	-	-	-	-	-	-
AfroLM	41.53	56.70	59.78	61.21	61.34	60.59	55.83	53.05	47.86	52.69	-	-	-	-	-	-	-	-	-	-	-	-	-	-
XLM-R large	24.50	38.48	45.12	45.59	48.24	47.11	54.42	55.94	59.18	62.56	62.85	63.87	63.78	66.25	61.35	62.97	64.06	61.56	66.63	63.88	64.50	57.93	60.62	60.02
XLM-R base	27.88	38.72	40.25	50.34	55.51	58.38	60.84	59.47	59.97	60.29	58.07	56.15	-	-	-	-	-	-	-	-	-	-	-	-
Nguni XLM-R	21.82	37.90	41.56	39.48	41.09	45.24	46.61	48.46	48.59	50.68	52.65	53.80	56.95	56.55	57.94	55.78	56.72	54.92	58.98	60.57	57.51	54.14	53.79	53.21

Table 21: NER results for Hausa with baselines. F1-gains reported.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Afro-XLMR large	18.06	33.16	42.93	45.43	46.70	52.71	58.34	57.68	61.75	61.55	74.87	67.68	70.34	71.10	68.55	68.43	76.27	70.60	71.99	70.98	70.16	65.99	74.51	72.33
Afro-XLMR base	18.99	32.69	38.39	43.66	48.29	53.01	53.33	69.52	58.42	58.94	58.97	59.71	-	-	-	-	-	-	-	-	-	-	-	-
AfriBERTa	23.74	37.15	43.93	49.38	53.58	53.61	56.02	57.80	55.29	57.54	-	-	-	-	-	-	-	-	-	-	-	-	-	-
AfroLM	25.84	41.47	44.29	45.61	47.65	47.97	45.21	40.42	43.84	44.88	-	-	-	-	-	-	-	-	-	-	-	-	-	-
XLM-R large	17.09	35.30	44.30	39.39	48.20	50.08	53.03	61.60	63.17	59.99	67.71	62.88	74.20	72.01	67.68	71.11	64.22	65.63	65.69	71.23	70.44	62.20	70.72	62.09
XLM-R base	22.75	34.33	37.45	42.11	47.30	47.79	50.55	52.60	56.31	50.12	54.01	58.46	-	-	-	-	-	-	-	-	-	-	-	-
Nguni XLM-R	16.53	41.49	44.33	53.59	40.85	54.25	44.29	58.83	63.53	61.18	65.76	74.53	65.59	65.73	68.71	67.68	68.71	63.84	66.70	71.44	75.22	73.13	67.08	72.13

Table 22: NER results for Naija Pidgin with baselines. F1-gains reported.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Afro-XLMR large	12.52	44.57	48.88	61.47	63.59	53.76	54.08	57.94	62.18	62.66	67.74	69.82	69.10	67.66	68.83	67.92	65.19	71.86	77.96	64.11	76.26	71.37	74.75	73.88
Afro-XLMR base	22.25	44.62	48.70	54.35	59.01	63.01	60.87	65.27	62.99	64.71	62.24	62.62	-	-	-	-	-	-	-	-	-	-	-	-
AfriBERTa	30.47	33.57	55.04	63.38	57.23	60.61	61.07	62.88	59.52	67.36	-	-	-	-	-	-	-	-	-	-	-	-	-	-
AfroLM	25.55	37.62	50.94	54.02	61.19	60.32	59.44	62.07	58.77	57.87	-	-	-	-	-	-	-	-	-	-	-	-	-	-
XLM-R large	26.16	48.60	48.00	50.96	50.90	50.95	51.86	63.43	55.78	57.33	63.72	60.22	62.44	62.26	60.65	58.27	64.62	58.00	60.07	68.97	68.43	65.55	67.21	65.29
XLM-R base	25.99	46.00	50.48	49.54	49.50	53.72	51.33	55.04	55.04	55.16	54.15	49.77	-	-	-	-	-	-	-	-	-	-	-	-
Nguni XLM-R	10.92	53.30	52.89	58.55	57.53	56.51	61.44	61.74	62.29	67.18	68.59	71.33	68.73	70.44	70.83	68.17	79.57	65.84	69.96	80.18	71.51	72.52	78.34	76.78

Table 23: NER results for isiXhosa with baselines. F1-gains reported.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Afro-XLMR large	9.86	25.32	30.56	32.84	35.94	34.24	38.24	41.45	46.00	46.00	48.78	49.10	52.89	53.30	54.53	53.75	55.20	51.53	53.72	52.62	53.12	53.25	54.05	50.78
Afro-XLMR base	10.00	24.91	30.63	39.38	43.26	48.50	49.22	50.55	49.45	48.66	47.63	-	-	-	-	-	-	-	-	-	-	-	-	-
AfriBERTa	25.18	32.69	40.97	47.67	48.69	50.43	50.39	48.54	52.54	53.78	-	-	-	-	-	-	-	-	-	-	-	-	-	-
AfroLM	21.85	29.26	33.68	39.90	44.07	44.87	43.92	44.49	46.33	45.32	-	-	-	-	-	-	-	-	-	-	-	-	-	-
XLM-R large	10.31	27.02	25.92	29.50	35.01	32.63	36.21	40.57	41.47	41.43	41.36	40.88	43.71	44.35	44.00	45.03	41.32	44.52	43.45	42.23	43.58	43.67	42.73	41.62
XLM-R base	11.34	21.13	29.55	32.61	31.99	36.15	37.03	35.52	37.09	37.79	35.82	35.99	-	-	-	-	-	-	-	-	-	-	-	-
Nguni XLM-R	11.03	26.02	29.45	29.33	32.43	29.92	37.95	38.48	38.58	41.38	41.54	43.23	44.22	44.42	42.81	43.08	45.58	45.19	46.57	46.26	46.93	47.52	46.88	45.37

Table 24: NER results for Igbo with baselines. F1-gains reported.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Afro-XLMR large	66.18	74.16	79.83	81.30	80.67	79.62	80.46	81.09	82.35	82.35	82.14	80.46	81.72	79.62	81.72	80.88	84.24	82.77	80.67	65.13	63.24	68.70	77.52	80.67
Afro-XLMR base	70.80	70.80	71.85	74.16	76.47	71.22	72.48	75.00	79.41	80.46	76.47	75.42	-	-	-	-	-	-	-	-	-	-	-	-
AfriBERTa	76.05	80.88	80.25	80.25	81.09	80.25	80.46	73.32	82.56	82.35	-	-	-	-	-	-	-	-	-	-	-	-	-	-
AfroLM	63.03	66.18	72.48	75.00	69.96	72.06	76.05	74.79	72.27	75.42	-	-	-	-	-	-	-	-	-	-	-	-	-	-
XLM-R large	64.08	77.52	77.31	76.26	79.20	77.94	78.78	79.83	78.99	78.57	78.99	79.20	78.57	78.36	81.93	79.62	78.15	79.83	79.83	78.99	77.94	81.51	81.30	77.73
XLM-R base	69.54	72.27	70.59	65.76	70.59	69.12	70.17	71.85	71.22	74.58	75.84	84.03	-	-	-	-	-	-	-	-	-	-	-	-
Nguni XLM-R	72.48	67.23	77.94	75.84	78.78	77.73	76.47	76.05	75.21	78.15	80.25	79.62	78.99	78.99	77.10	80.04	80.04	78.78	76.68	71.01	68.49	71.85	75.00	78.78

Table 25: NTC results for Swahili. Accuracy reported.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Afro-XLMR large	39.09	19.09	41.82	64.55	60.91	56.36	62.73	33.64	64.55	60.91	68.18	65.45	71.82	64.55	70.91	70.91	73.64	74.55	60.00	19.09	40.91	42.73	46.36	53.64
Afro-XLMR base	60.00	45.45	60.00	42.73	60.91	44.55	68.18	64.55	62.73	45.45	70.00	50.00	-	-	-	-	-	-	-	-	-	-	-	-
AfriBERTa	50.91	9.09	68.18	65.45	60.91	55.45	60.91	10.91	60.91	74.55	-	-	-	-	-	-	-	-	-	-	-	-	-	-
AfroLM	60.91	50.00	63.64	68.18	6.36	63.64	12.73	16.36	63.64	69.09	-	-	-	-	-	-	-	-	-	-	-	-	-	-
XLM-R large	61.82	51.82	49.09	48.18	49.09	46.36	50.00	40.00	36.36	50.91	41.82	46.36	50.91	53.64	45.45	50.91	41.82	47.27	52.73	54.55	15.45	57.27	39.09	50.91
XLM-R base	21.82	54.55	50.91	50.91	46.36	64.55	50.91	50.00	51.82	50.91	42.73	60.91	-	-	-	-	-	-	-	-	-	-	-	-
Nguni XLM-R	60.00	60.91	65.45	61.82	65.45	52.73	58.18	59.09	66.36	67.27	71.82	71.82	72.73	73.64	73.64	77.27	74.55	66.36	61.82	57.27	45.45	45.45	30.91	53.64

Table 26: NTC results for Luganda. Accuracy reported.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Afro-XLMR large	75.39	79.81	82.02	82.97	79.50	81.70	80.76	82.02	85.80	86.44	88.33	87.07	85.80	85.49	83.91	85.17	85.17	85.49	84.54	77.29	70.03	73.19	81.70	84.23
Afro-XLMR base	80.76	74.45	77.29	76.03	79.50	82.02	81.07	82.33	86.12	84.86	85.17	80.44	-	-	-	-	-	-	-	-	-	-	-	-
AfriBERTa	84.23	84.54	84.23	86.44	84.54	84.54	83.28	85.49	87.38	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
AfroLM	81.70	82.65	84.54	85.80	80.44	80.76	82.33	81.07	80.44	84.86	-	-	-	-	-	-	-	-	-	-	-	-	-	-
XLM-R large	77.92	79.81	81.07	81.39	78.86	80.13	82.65	78.55	81.39	83.28	83.60	82.02	85.49	84.23	87.07	85.49	84.54	84.54	85.49	85.80	86.44	85.17	87.38	84.54
XLM-R base	76.03	73.19	72.24	73.82	77.29	71.92	76.34	77.92	75.39	78.55	80.13	87.38	-	-	-	-	-	-	-	-	-	-	-	-
Nguni XLM-R	81.70	73.82	82.97	79.18	80.13	81.07	83.28	81.07	81.70	81.39	82.65	79.18	80.44	78.55	80.76	82.65	82.33	82.02	78.23	70.03	70.66	71.61	79.18	78.23

Table 27: NTC results for Hausa. Accuracy reported.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Afro-XLMR large	82.89	82.89	90.13	88.16	89.47	88.82	88.16	91.45	89.47	88.82	90.13	87.50	90.13	87.50	90.13	89.47	92.11	88.82	88.82	82.89	80.92	84.87	90.13	90.79
Afro-XLMR base	88.16	83.55	86.18	84.87	90.13	86.18	88.16	88.16	89.47	90.79	89.47	88.16	-	-	-	-	-	-	-	-	-	-	-	-
AfriBERTa	89.47	91.45	92.11	92.76	91.45	90.79	90.79	88.82	91.45	90.13	-	-	-	-	-	-	-	-	-	-	-	-	-	-
AfroLM	88.82	88.82	88.82	87.50	88.82	89.47	85.53	90.79	91.45	89.47	-	-	-	-	-	-	-	-	-	-	-	-	-	-
XLM-R large	83.55	92.76	91.45	90.13	90.13	92.76	90.13	88.82	90.13	88.16	90.13	87.50	88.16	88.16	90.13	88.16	88.82	88.16	90.13	90.79	88.16	86.18	88.16	90.79
XLM-R base	84.21	89.47	88.16	84.87	88.16	87.50	83.55	78.29	82.89	88.16	82.24	92.11	-	-	-	-	-	-	-	-	-	-	-	-
Nguni XLM-R	88.16	82.89	90.13	89.47	92.76	90.79	90.13	91.45	89.47	89.47	90.79	90.13	90.13	92.11	90.13	89.47	90.79	88.82	90.13	86.84	82.89	84.21	92.11	92.76

Table 28: NTC results for Naija Pidgin. Accuracy reported.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Afro-XLMR large	72.39	75.76	78.79	77.78	79.46	81.14	80.81	82.49	87.54	86.53	89.56	87.21	88.89	92.93	89.23	89.90	89.23	89.56	85.86	77.78	73.06	82.83	84.18	88.55
Afro-XLMR base	78.45	78.11	76.43	77.78	84.51	83.50	86.20	86.20	87.54	84.18	86.20	83.16	-	-	-	-	-	-	-	-	-	-	-	-
AfriBERTa	77.78	80.81	79.12	81.48	80.13	76.77	78.79	79.46	80.81	79.46	-	-	-	-	-	-	-	-	-	-	-	-	-	-
AfroLM	75.08	75.76	78.79	78.45	79.46	80.47	77.10	76.77	79.46	77.44	-	-	-	-	-	-	-	-	-	-	-	-	-	-
XLM-R large	78.45	78.11	79.46	79.46	79.12	76.77	73.74	77.10	76.77	78.45	79.46	81.48	81.14	81.48	81.82	81.82	83.16	84.18	86.87	86.53	85.52	85.86	82.49	82.15
XLM-R base	71.72	73.06	71.72	75.08	74.41	76.77	77.10	76.43	80.13	77.44	74.75	82.15	-	-	-	-	-	-	-	-	-	-	-	-
Nguni XLM-R	75.42	73.06	87.88	85.86	87.88	87.54	86.53	89.56	94.61	91.25	91.25	90.24	87.88	92.59	93.94	92.93	93.60	93.60	89.90	81.14	79.46	80.81	87.54	90.24

Table 29: NTC results for isiXhosa. Accuracy reported.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Afro-XLMR large	57.73	64.43	75.26	74.74	72.68	70.62	79.38	80.93	77.32	77.32	77.84	79.38	78.35	80.41	79.90	82.47	79.90	76.80	76.80	70.62	67.01	68.56	77.84	79.38
Afro-XLMR base	65.98	59.28	69.59	63.92	66.49	70.10	70.62	72.16	75.26	76.80	74.23	73.71	-	-	-	-	-	-	-	-	-	-	-	-
AfriBERTa	74.23	80.81	78.37	78.35	77.84	77.84	78.87	76.29	82.96	84.02	-	-	-	-	-	-	-	-	-	-	-	-	-	-
AfroLM	67.01	66.49	71.13	73.20	75.26	70.62	67.01	74.23	76.29	75.77	-	-	-	-	-	-	-	-	-	-	-	-	-	-
XLM-R large	69.07	69.07	73.20	69.07	68.04	66.49	68.56	69.07	68.56	70.62	68.56	67.53	65.98	62.89	67.01	68.04	70.10	69.59	71.65	68.56	71.13	67.01	71.13	73.20
XLM-R base	64.43	63.92	70.10	64.95	63.92	59.79	62.37	68.04	64.43	65.46	63.92	74.23	-	-	-	-	-	-	-	-	-	-	-	-
Nguni XLM-R	67.01	69.59	73.20	78.87	69.59	75.26	72.16	77.32	79.38	79.90	82.47	80.93	78.87	81.44	81.96	81.96	81.44	80.93	72.68	61.34	62.89	65.46	71.65	76.29

Table 30: NTC results for Igbo. Accuracy reported.