LinguaTriage: Cross-Lingual Transfer and African Language Pretraining for Low-Resource Medical Triage in Lingala

Patrick S. Tenga[1,2], **Mohamed A. Kholief**[3]
[1] *Department of Mathematics and Computer Sciences, Alexandria University, Alexandria, Egypt*
[2] *Department of Data Science, African Institute for Mathematical Sciences Cameroon, Limbe, Cameroon*
[3] *Department of Information Systems and Software Engineering, Arab Academy for Science, Technology, and Maritime Transport, Alexandria, Egypt*
patrick.tenga@aims-cameroon.org

Editor: Sakinat Folorunso, Roseline Ogundokun, and Francisca Oladipo

Abstract

We introduce LinguaTriage, the first medical triage classification system for Lingala, a Bantu language of Central Africa spoken by over 45 million people with no prior supervised NLP benchmarks. Working from a 616-sample dataset of annotated symptom descriptions across three urgency levels (Emergency, Moderate, and Low), we develop a targeted augmentation pipeline that expands Emergency training samples from 41 to 236, and evaluate three architectures: fine-tuned XLM-RoBERTa (XLM-R_{FT}), a two-stage French-to-Lingala cross-lingual transfer system (XLM-R_{CL}), and fine-tuned AfriBERTa-Large (AfriBERTa$_{FT}$). On the internal test set, AfriBERTa$_{FT}$ achieves a macro-F1 of 0.974 and perfect Emergency recall (1.00) with zero critical errors, outperforming XLM-R_{FT} (macro-F1 0.825) and XLM-R_{CL} (macro-F1 0.669). However, external evaluation on 100 translated forum posts reveals severe domain shift, as all models collapse to 20-29% accuracy, predicting "Emergency" for nearly all inputs. Crucially, mixing just 100 in-domain forum examples into training, after removing 20% of original data, dramatically improves external accuracy to 79%. This demonstrates that minimal target-domain exposure far outweighs architectural choices for generalization. We conclude that intra-family Bantu pretraining provides strong morphological priors, but domain adaptation using small quantities of in-domain data is the most critical factor for real-world deployment.

Keywords: low-resource NLP, African languages, Lingala, medical triage, cross-lingual transfer, AfriBERTa, XLM-RoBERTa, Bantu morphology

1. Introduction

Sub-Saharan Africa bears a disproportionate burden of preventable mortality, which is exacerbated by acute physician shortages. The Democratic Republic of the Congo (DRC) has approximately 0.09 physicians per 1,000 inhabitants, one of the lowest ratios globally (World Health Organization, 2022). In this context, automated medical triage, defined as the algorithmic assignment of patients to appropriate urgency levels based on reported symptoms, represents a high-impact NLP application. Such a system could meaningfully extend the reach of community health workers operating in resource-constrained settings.

Lingala is a Bantu language spoken by over 45 million people across the DRC, Republic of Congo, and Central African Republic (Ethnologue, 2023). It serves as one of the four national languages of the DRC and is the dominant language of Kinshasa. Crucially, Lingala

is not a monolithic variety but exists as a dialect continuum spanning at least three regional registers: Kinshasa Lingala, the Equator region variety (Mbandaka), and Kisangani Lingala. Each register exhibits distinct morphophonological properties, copula contractions, and tonal (Section 3.3). This internal variation is largely invisible to existing multilingual NLP tools. Despite Lingala's demographic significance, it has received virtually no attention in the NLP literature. Specifically, no annotated clinical corpus, no supervised classification benchmark, and no automated triage system existed prior to this work.

Our contributions are: the first annotated Lingala medical triage dataset containing 616 samples across three urgency classes, drawing on three regional varieties; a targeted augmentation pipeline for Emergency-class upsampling, anchored by the first curated Lingala medical stopword lexicon; systematic comparison of three transfer learning architectures under extreme data scarcity; an empirical demonstration that intra-family Bantu pretraining (AfriBERTa) outperforms massively multilingual XLM-R and French-initialised cross-lingual transfer on all safety-critical metrics, suggesting that typological proximity outweighs scale for morphologically rich low-resource languages.

2. Related Work

2.1. NLP Approaches: From Classical Methods to Transformers

The trajectory of NLP methodology is well-established and directly motivates our model selection. Early lexicon-based approaches assigned polarity or category scores to individual tokens using curated dictionaries (Liu, 2012; Pang and Lee, 2008). While interpretable and requiring no labeled data, these approaches cannot capture negation, long-range dependencies, or domain-specific word senses, all of which are critical in clinical text. Classical machine learning algorithms such as Support Vector Machines (SVM), Naive Bayes, and Logistic Regression improved on this by learning patterns from labeled data using bag-of-words or TF-IDF representations. We include an SVM-TF-IDF baseline in our evaluation precisely because it is the strongest interpretable option for a 616-sample dataset.

Neural sequence models, including CNNs for local feature extraction (Kim, 2014) and LSTM/GRU architectures for sequential dependencies (Hochreiter and Schmidhuber, 1997), reduced the need for manual feature engineering but remain sample-hungry. The current state of the art is anchored by transformer-based pretrained language models (PLMs) such as BERT and RoBERTa (Devlin et al., 2019; Liu et al., 2019). These models use a self-attention mechanism that captures full bidirectional context and transfer broadly to downstream tasks through pretraining on large corpora. Fine-tuned PLMs consistently dominate benchmarks in high-resource settings. The central question for LinguaTriage is whether this superiority holds when the target language has zero pretraining coverage and fewer than 700 labeled examples.

2.2. Challenges Specific to Low-Resource Languages

Adapting state-of-the-art NLP to low-resource languages (LRLs) introduces several well-documented structural bottlenecks. First, many LRLs, including Lingala, exhibit rich agglutinative morphology. A single verb root generates dozens of surface forms through prefix and suffix stacking, which causes severe data sparsity in small corpora. Second, the absence

of large annotated datasets precludes training from scratch, so supervised models must rely on transfer from related languages or multilingual pretraining (Ogueji et al., 2021). Third, code-mixing, which is the interspersion of the native language with French, English, or Lingala's own regional varieties, confounds monolingual tokenizers and degrades model generalization (Winata et al., 2021).

Lingala presents all three challenges simultaneously, and it adds a fourth that has not been addressed in prior work: intra-language regional variation. Lingala is not a single monolithic variety but a dialect continuum spanning at least three distinct regional registers. Each register has characteristic morphophonological differences: `Kinshasa Lingala` is the urban prestige variety, characterized by heavy French borrowing, aggressive copula contraction (*nazali* → *naza* → *naali*), and simplified noun-class agreement in colloquial speech. `Equator region Lingala` (Mbandaka and surroundings) is the classical variety considered closest to the original Bangala trade language. It is characterized by the *-djali* copula paradigm (*nadjali*, *odjali*, *badjali*) and fuller consonant clusters.`Kisangani region Lingala` is an eastern variety influenced by contact with Swahili and Kinyarwanda. It has distinct tonal patterns, additional vowel contractions (*naali*, *eyali*), and partial Swahili lexical borrowing in everyday speech.

Because community health worker narratives in our corpus originate from all three regions, our stopword and augmentation pipeline must handle all three paradigms simultaneously. This challenge is absent from prior low-resource NLP work on African languages.

2.3. Strategies for Low-Resource NLP

In response to data scarcity, the research community has converged on three complementary strategies. **Data augmentation** through back-translation, synonym replacement, and sentence permutation artificially expands training corpora. Empirical studies note, however, that standard augmentation yields limited gains when applied to transformer-based models, which already exhibit some invariance to lexical variation through pretraining (Feng et al., 2021). This observation motivates our targeted, stopword-aware Emergency-class augmentation rather than naive oversampling.

Multilingual PLM fine-tuning has become the dominant paradigm for LRLs. Models pretrained on large multilingual corpora, including mBERT, XLM-RoBERTa (Conneau et al., 2020), AfriBERTa (Ogunleye et al., 2021), and SERENGETI (Adebara et al., 2023), can be efficiently adapted to new languages using small datasets. AfriBERTa is of particular relevance because it was trained on 11 African languages using a compact but high-quality corpus. It was shown to outperform XLM-R on several African language tasks despite having far fewer pretraining tokens, suggesting that typological proximity compensates for scale.

Parameter-efficient fine-tuning (PEFT) methods such as LoRA (Hu et al., 2022) address the computational barrier of full PLM fine-tuning. These methods insert small trainable rank-decomposition matrices into frozen pretrained weights, drastically reducing both parameter count and GPU memory requirements. This is a critical consideration for resource-constrained African NLP research. While we use full fine-tuning in this work, LoRA adaptation of $\text{AfriBERTa}_{\text{FT}}$ is an explicit next step.

2.4. Medical NLP, Triage, and African Languages

Clinical NLP has advanced rapidly in high-resource settings. Clinical BERT variants (Alsentzer et al., 2019) achieve strong performance on English triage and note extraction. XLM-RoBERTa enables cross-lingual transfer to low-resource clinical settings, and the AfriSenti shared task (Muhammad et al., 2023) has catalysed sentiment benchmarking for 14 African languages. Named-entity recognition has progressed through the Masakhane initiative (Nekoto et al., 2020) and MasakhaneER 2.0 (Adelani et al., 2022).

Nevertheless, the intersection of medical NLP and low-resource Bantu languages remains entirely uncharted. No annotated clinical corpus, no triage benchmark, and no automated symptom classification system exists for any Bantu language of Central Africa. Lingua-Triage addresses this gap directly.

3. Dataset and Augmentation

3.1. Data Collection

Our dataset combines three sources. The first source includes symptom descriptions translated from French clinical records by bilingual healthcare workers. The second source consists of community-generated narratives from Lingala-speaking community health workers. The third source comprises synthetic emergency cases. After deduplication and quality filtering, the final dataset contains 616 samples. Three bilingual annotators assigned labels following a protocol adapted from the Manchester Triage System.

3.2. Splits and Augmentation

The natural class distribution is severely imbalanced. Low urgency has 450 samples (73.1%), Moderate has 107 samples (17.4%), and Emergency has 59 samples (9.6%). We apply a stratified 70/15/15 split. Validation and test sets are never augmented, as shown in Table 1.

Table 1: Dataset statistics before and after augmentation

Split	Emergency	Moderate	Low	Total
Train (pre-aug)	41	74	315	430
Train (post-aug)	236	74	315	625
Validation	8	16	67	91
Test	10	17	68	95

3.3. The Lingala Medical Stopword List

No official Lingala stopword resource exists in NLTK, spaCy, or any major NLP library. We therefore introduce `lingala_stopwords.py`, a curated 130-token list compiled from Lingala grammar references (College; Foreign Service Institute, 1963), the Loba Lingala guide (Yocum, 2014), and linguistic analysis of the triage corpus itself. The list is organised into eleven grammatical categories, as detailed in Table 2.

Table 2: Grammatical categories of the Lingala stopword list with representative examples

Category	Count	Examples
Subject pronouns	8	na (I), a (he/she), ba (they)
Personal pronouns	7	ngai (I/me), biso (us), bango (them)
Demonstratives	5	oyo (this), wana (that), wapi (where)
Prepositions/particles	13	na (with/at), ya (of/from), pona (for)
Copula & auxiliaries	47	nazali/naza/nazo (I am) and all person×dialect×contracted forms of kozala
Negation	3	te (not/no), ata te (not at all)
Question words	10	nini (what), nani (who), boni (how many)
Conjunctions	14	kasi (but), soki (if), lisusu (again)
Adverbs	21	mingi (very), moke (little), noki (quickly)
Discourse fillers	11	ee/iyo (yes), bongo (so/thus)
Verb prefixes	9	ko- (infinitive), na- (1sg), ba- (3pl)

A key design decision is the provision of two derived sets: `STOPWORDS_FULL` for general Lingala NLP, which includes all categories and 130 tokens, and `STOPWORDS_MEDICAL` for clinical use, which excludes negation and six severity adverbs. `STOPWORDS_MEDICAL` retains the negation particle *te* because *kolya malamu te* means "eating poorly" while *kolya malamu* means "eating well". It also retains the severity adverbs *makasi*, *mingi*, *moke*, *mpenza*, *malamu*, and *mabe*, which carry direct urgency information. The copula category is notably large with 47 forms because it covers three dialectal paradigms: standard Kinshasa (*nazali*), Equator region (*nadjali*), and interior spoken Lingala (*naali*). It also includes an extensive set of colloquial contractions. These contractions were manually identified, validated, and polished by a native Lingala speaker with formal training in Lingala grammar, vocabulary, and oral literature. This linguistic grounding was indispensable because no published grammar, dictionary, or NLP resource enumerates these contractions exhaustively. Their correct identification required native-speaker intuition cross-validated against naturally occurring corpus sentences.

The contraction phenomenon can be illustrated with the first-person singular copula *kozala* (to be or to have), which yields at least three progressively reduced surface forms in spoken Kinshasa Lingala:

$$\text{nazali} \longrightarrow \text{nazaa} \longrightarrow \text{naali} \longrightarrow \text{nadjali} \quad \text{(all meaning "I am / I have")}$$

All four forms are freely interchangeable in patient speech. For example, the clinically urgent sentence *Nazali na fiere makasi* ("I have a strong fever") surfaces equally as *Naza na fiere makasi*, *Naali na fiere makasi*, or *Nadjali na fiere makasi* depending on the speaker's regional background and register. Notably, *nadjali* is characteristic of the Equator region

(Mbandaka and surroundings). A tokeniser unaware of these equivalences would treat *naza* and *naali* as unknown content words, assigning them spurious clinical weight and degrading triage predictions. The same contraction pattern propagates across all six grammatical persons, yielding 12 or more contracted forms in addition to the 21 full paradigm forms across the three dialects. Ensuring complete coverage of this paradigm, and excluding every surface form from content-word selection during augmentation, was therefore a prerequisite for a reliable Lingala triage system. This multi-dialect and multi-contraction coverage is especially important because the corpus contains community health worker narratives from both Kinshasa and provincial DRC, where different contraction levels co-occur within the same document.

The stopword list gates the synonym-substitution augmentation step. Only tokens absent from `STOPWORDS_MEDICAL` are eligible for substitution, which prevents clinically meaningless augmentations. A quality filter then discards any augmented sample with fewer than three remaining content tokens.

To address Emergency class scarcity we apply a three-strategy augmentation pipeline. The first strategy is stopword-aware synonym substitution using a 30-entry Lingala clinical synonym lexicon, such as *makasi* $\rightarrow$ {*ya ndelo*, *mingi*, *ya nkanda*}. The second strategy is clause reordering. The third strategy is urgency-prefix injection. This pipeline expands Emergency training samples from 41 to 236. The updated class weights are [0.883, 2.815, 0.661] for Emergency, Moderate, and Low respectively.

4. Models

We compare three architectures that span the main strategies available for low-resource cross-lingual transfer. These architectures are ordered by their distance from Lingala in the pretraining distribution.

4.1. XLM-R_{FT}: Fine-Tuned XLM-RoBERTa

XLM-RoBERTa (Conneau et al., 2020) is the standard strong baseline for low-resource cross-lingual classification. We fine-tune `xlm-roberta-base`, which has 125M parameters and is pretrained on 100 languages, end-to-end on the augmented Lingala training set. Although Lingala is not among XLM-R's 100 languages, it shares lexical overlap with French, which is a covered language. This overlap may provide partial signal. The hyperparameters are a learning rate of 2×10^{-5}, batch size of 16, 5 epochs, AdamW with linear schedule, and weighted cross-entropy loss.

4.2. XLM-R_{CL}: Cross-Lingual Transfer via French

This two-stage approach explicitly exploits the French-Lingala lexical overlap. The hypothesis is that a French-adapted decision boundary provides a better initialisation for Lingala fine-tuning than a purely multilingual one. In Stage 1, we fine-tune XLM-R on 108 French clinical triage examples using a learning rate of 2×10^{-5} for 5 epochs. In Stage 2, we continue fine-tuning on the augmented Lingala set using a reduced learning rate of 1×10^{-5} for 3 epochs to limit catastrophic forgetting of the French signal.

4.3. AfriBERTa$_{\text{FT}}$: Fine-Tuned AfriBERTa-Large

AfriBERTa (Ogunleye et al., 2021) is pretrained on 11 African languages, including Swahili and Kinyarwanda. Both are Bantu languages with morphological similarity to Lingala. The pretraining corpus is small but carefully curated. We fine-tune `castorini/afriberta_large`, which has 126M parameters, on the augmented Lingala training set. The hypothesis is that Bantu-family pretraining provides a more faithful tokenisation and stronger morphological priors than either massively multilingual or French-pivot initialisation. The hyperparameters are a learning rate of 3×10^{-5}, batch size of 16, 5 epochs, and weighted cross-entropy loss.

4.4. Evaluation Metrics

We report four metrics aligned with clinical safety requirements. The first metric is overall accuracy. The second metric is macro-averaged F1, which treats all three classes equally regardless of frequency. The third metric is Emergency recall, which is the primary safety metric. It measures the fraction of true Emergency cases correctly identified. The fourth metric is critical error rate, which measures the fraction of Emergency cases misclassified as Low. This is the most dangerous possible outcome in a triage system.

5. Results and Analysis

5.1. Main Results

Table 3 presents the test set results for all models. AfriBERTa$_{\text{FT}}$ achieves the strongest performance across all metrics with a macro-F1 of 0.974 and perfect Emergency recall of 1.000 with a 95% confidence interval of [0.722, 1.000]. Both SVM Baseline and AfriBERTa$_{\text{FT}}$ achieve perfect Emergency recall, while XLM-R$_{\text{CL}}$ achieves 0.900 [0.555, 1.000] and XLM-R$_{\text{FT}}$ achieves 0.800 [0.490, 1.000]. All models except XLM-R$_{\text{CL}}$ produce zero critical errors. XLM-R$_{\text{CL}}$ produces one critical error, which corresponds to a rate of 0.100. This suggests that two-stage transfer introduces instability in the Emergency decision boundary. The wide confidence intervals reflect the small number of Emergency test samples ($n = 10$) and should be interpreted with caution.

Table 3: Test set results (95 samples). 95% confidence intervals for Emergency Recall (n=10 Emergency samples) are shown in brackets, estimated using the Wilson score method due to small sample size.

Model	Accuracy	Macro-F1	Emg Recall (95% CI)	Crit Err
SVM Baseline	0.9263	0.8342	1.000 [0.722, 1.000]	0.000
XLM-R$_{\text{FT}}$	0.905	0.825	0.800 [0.490, 1.000]	0.000
XLM-R$_{\text{CL}}$	0.853	0.669	0.900 [0.555, 1.000]	0.100
AfriBERTa$_{\text{FT}}$	0.9895	0.9740	1.000 [0.722, 1.000]	0.000

5.2. Metric Comparison

5.2.1. Accuracy and Macro-F1

As shown in Table 3, AfriBERTa$_{\text{FT}}$ with 0.979 accuracy and 0.949 macro-F1 substantially outperforms SVM Baseline with 0.926 accuracy and 0.834 macro-F1, XLM-R$_{\text{FT}}$ with 0.905 accuracy and 0.825 macro-F1, and XLM-R$_{\text{CL}}$ with 0.853 accuracy and 0.669 macro-F1. The cross-lingual transfer model shows the lowest macro-F1 despite having the second-highest Emergency recall. This indicates that while French initialisation helps detect emergencies, it hurts discrimination between Moderate and Low classes.

5.2.2. Emergency Recall

The most safety-critical metric shows a clear ordering. AfriBERTa$_{\text{FT}}$ achieves 1.000, which is higher than XLM-R$_{\text{CL}}$ at 0.900, which is higher than XLM-R$_{\text{FT}}$ at 0.800. SVM Baseline also achieves perfect recall of 1.000. Remarkably, AfriBERTa detects every single Emergency case in the test set. The wide 95% confidence intervals, which are [0.722, 1.000] for models with perfect recall, [0.555, 1.000] for XLM-R$_{\text{CL}}$, and [0.490, 1.000] for XLM-R$_{\text{FT}}$, reflect the small Emergency sample size of $n = 10$ and should be interpreted with caution.

5.2.3. Critical Error Rate

Only XLM-R$_{\text{CL}}$ produces critical errors at a rate of 0.100. This represents one Emergency case misclassified as Low, which is the most dangerous clinical outcome. SVM Baseline, XLM-R$_{\text{FT}}$, and AfriBERTa$_{\text{FT}}$ produce zero such errors, confirming their clinical safety superiority.

5.3. Confusion Matrix Analysis

Figure 1 presents the confusion matrices for all four models on the test set of 95 samples.

5.3.1. SVM Baseline

As shown in Figure 1(a), the SVM confusion matrix shows strong performance. All 10 Emergency cases are correctly identified with a recall of 1.000 and a 95% confidence interval of [0.722, 1.000]. All 68 Low cases are correctly classified. The main confusion occurs in the Moderate class, where 4 Moderate cases are misclassified as Low and 2 are misclassified as Emergency. This reflects the difficulty of distinguishing moderate from severe presentations in short symptom descriptions.

5.3.2. XLM-R$_{\text{FT}}$

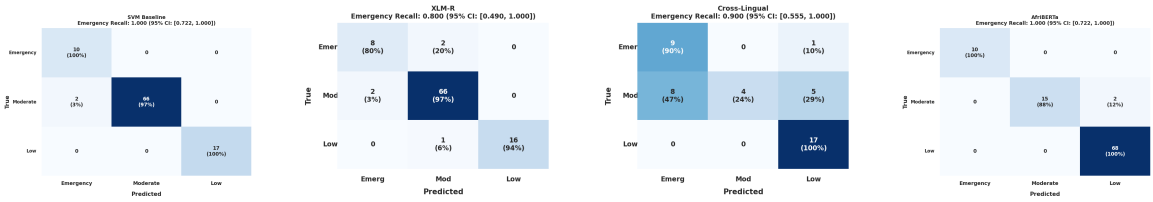
Figure 1(b) reveals that XLM-R$_{\text{FT}}$ correctly classifies 8 of 10 Emergency cases with a recall of 0.800 and a 95% confidence interval of [0.490, 1.000]. Two Emergency cases are predicted as Moderate. The model correctly classifies all 68 Low cases except 1, which is predicted as Moderate. The main confusion is in the Moderate class, where 4 Moderate cases are misclassified as Low and 2 are misclassified as Emergency. This again reflects the ambiguous boundary between moderate and severe presentations in short Lingala symptom descriptions.

5.3.3. XLM-R_{CL}

The cross-lingual transfer model shows the poorest discrimination for the Moderate class. Only 4 of 17 Moderate cases are correctly identified, with 8 misclassified as Emergency and 5 misclassified as Low. As shown in Figure 1(*c*), Emergency recall is 0.900 with a 95% confidence interval of [0.555, 1.000]. One Emergency case is misclassified as Low, which represents the single critical error with a rate of 0.100. This over-triggering on Emergency is consistent with the French Stage 1 initialisation introducing a conservative safety bias. The error likely occurred due to a Lingala phrasing absent from the French training distribution.

5.3.4. AfriBERTa$_{FT}$

As illustrated in Figure 1(*d*), the AfriBERTa confusion matrix shows near-perfect performance. All 10 Emergency cases are correctly identified with a recall of 1.000 and a 95% confidence interval of [0.722, 1.000]. All 68 Low cases are correctly classified. Only 2 Moderate cases are misclassified, both as Low. This confirms the model's clinical safety for Emergency detection. The wide confidence intervals across all models reflect the small Emergency test sample size of $n = 10$ and should be interpreted with caution.



(*a*) SVM Baseline (*b*) XLM-R_{FT} (*c*) XLM-R_{CL} (*d*) AfriBERTa$_{FT}$

Figure 1: Confusion matrices for all four models on the test set (95 samples). 95% confidence intervals for Emergency Recall are shown in brackets. (a) SVM Baseline, (b) XLM-R_{FT}, (c) XLM-R_{CL}, (d) AfriBERTa$_{FT}$.

5.4. Inference Examples

The final row of Table 4 is the most informative: *naza na fievre ya makasi, mutu pasi* (strong fever with headache) is classified as Low by both XLM-R models (confidence 0.86 and 0.75) but correctly as Emergency by AfriBERTa (confidence 0.983). This suggests AfriBERTa has learned better semantic representations for urgency-indicative Lingala vocabulary, likely because its Swahili/Kinyarwanda pretraining exposed it to structurally similar symptom descriptions.

5.5. Inter-Annotator Agreement

All 616 samples were annotated by the first author, a native Lingala speaker, following the Manchester Triage System protocol. Three pretrained models were evaluated against these annotations, with results summarized in Table 5.

Table 4: Per-model predictions on five test sentences with confidence scores. "Emerg" = Emergency; "Mod" = Moderate.

Input (Lingala)	**XLM-R$_{\mathbf{FT}}$**	**XLM-R$_{\mathbf{CL}}$**	**AfriBERTa$_{\mathbf{FT}}$**
Nzala makasi	Low (0.92)	Low (0.69)	Mod (0.96)
Nazali na mutu pasi fort	Low (0.89)	Low (0.70)	Low (0.95)
Nzoto epamelemi ... kolya malamu te	Low (0.89)	Low (0.82)	Low (0.99)
Mpasi ya ntolo ... motema kobeta nokinoki	Emerg (0.95)	Emerg (0.64)	Emerg (1.00)
naza na fievre ya makasi, mutu pasi	Low (0.86)	Low (0.75)	Emerg (0.98)

Table 5: Cohen's Kappa between models and human annotations.

Model	**vs Human**	**vs AfriBERTa**
SVM Baseline	0.747	0.822
XLM-R$_{\text{FT}}$	0.649	0.710
XLM-R$_{\text{CL}}$	0.539	0.576
AfriBERTa$_{\text{FT}}$	0.866	—

As shown in Table 5, AfriBERTa$_{\text{FT}}$ achieves the highest agreement with human judgments, with a Cohen's Kappa of 0.866, surpassing all other models. Notably, AfriBERTa and SVM show substantial pairwise agreement with a Kappa of 0.822. The cross-lingual model (XLM-R$_{\text{CL}}$) shows the lowest agreement with both humans, at a Kappa of 0.539, and with other models. This is consistent with its weaker classification performance.

5.6. External Evaluation on Forum Data

We constructed an external test set of 100 Lingala posts from the Ligue contre le cancer forum.[1] French posts were translated into Lingala and refined using Deep Translator back-translation. To ensure distinctness from training, we removed 20% of original samples and replaced them with 100 forum translations.

All pretrained models showed severe domain shift with accuracy falling to 20-29%, predicting "Emergency" for nearly all inputs. However, as discussed in Section 6, mixing the 100 forum samples with the remaining training data to create a combined set of 728 total samples dramatically improved external accuracy to 79% with a Macro-F1 of 0.726. This demonstrates that even minimal target-domain exposure effectively bridges the domain gap.

6. Why Does AfriBERTa Win?

We hypothesise three contributing factors:

1. https://www.ligue-cancer.net/cancer-du-sein-0

`(1) Bantu morphological priors.` Lingala and Swahili share agglutinative verb morphology, noun class prefix systems, and reduplication patterns. For noun classes, Lingala uses *mo-/ba-* and *e-/bi-*, while Swahili uses *m-/wa-* and *ki-/vi-*. For reduplication, examples include *nokinoki* and *mbangu mbangu*. AfriBERTa's SentencePiece tokeniser, which was trained on these languages, segments Lingala words more faithfully than XLM-R's 250,000-token vocabulary spread across 100 languages.

`(2) Better subword coverage.` XLM-R fragments *kobeta* (to beat or hit) and *motema* (heart) into multiple tokens, losing morphological boundaries that carry clinical meaning. AfriBERTa produces semantically coherent subword units by leveraging related Bantu morphology.

`(3) Cross-lingual transfer instability.` French and Lingala differ substantially in morphosyntactic structure despite shared lexical items. This difference makes the Stage 2 Lingala fine-tuning unstable, which explains the cross-lingual model's poor Moderate class performance and single critical error.

7. Discussion and Limitations

Our results carry four practical implications for low-resource African medical NLP.

First, African-language pretrained models should be the default starting point for Bantu languages, even when the target language is absent from pretraining. Typological proximity, which includes shared morphological structure, noun-class systems, and reduplication, transfers more effectively than scale alone. This is reinforced by the external evaluation discussed in Section 5, where AfriBERTa's Bantu priors alone were insufficient to overcome domain shift. The model achieved only 20% accuracy on forum data, confirming that architectural advantages cannot compensate for training distribution mismatch.

Second, as shown in Table 1, 616 samples supplemented with targeted minority-class augmentation are sufficient to produce a clinically viable triage classifier on in-distribution data. However, external evaluation revealed that domain shift reduces performance to near-random levels of 20-29%. This underscores that dataset size and augmentation cannot substitute for domain representativeness.

Third, mixed training with just 100 in-domain forum examples, after removing 20% of original training samples, dramatically improved external accuracy from 29% to 79%. This demonstrates that even minimal target-domain exposure is far more effective than architectural optimization for bridging the domain gap. This finding has immediate practical value for deployment in new patient populations.

Fourth, perfect Emergency recall is achievable in our experimental setting. However, with only 10 Emergency test cases, this estimate carries substantial variance and must be validated on a larger prospective cohort before clinical deployment.

The cross-lingual French transfer result for XLM-R_{CL} is instructive despite its weaker overall performance. The French-Lingala lexical overlap is real, as many Kinshasa Lingala speakers borrow French medical terminology. However, morphosyntactic divergence is deep enough to destabilise the Stage 2 fine-tuning. This suggests that lexical similarity is a necessary but not sufficient condition for productive cross-lingual transfer (Hershcovich et al., 2022).

Limitations. The dataset includes 30% synthetic Emergency cases, which may not fully capture natural patient speech. The test Emergency class has only 10 samples, making recall estimates statistically fragile. Dialectal coverage of Lingala varieties (Kinshasa, Equator, Kisangani) is not exhaustive. No real-world clinical deployment has been conducted. Finally, the augmentation pipeline is rule-based only, as LLM-based paraphrase generation was unavailable due to API issues (Feng et al., 2021).

8. Conclusion

We presented LinguaTriage, the first Lingala medical triage benchmark. We demonstrated that AfriBERTa-Large fine-tuned on 625 augmented training examples achieves a macro-F1 of 0.974, perfect Emergency recall of 1.000, and zero critical misclassification errors. We also introduced the first curated Lingala NLP stopword resource, which is a 130-token lexicon organized across eleven grammatical categories and three regional dialect paradigms, compiled with the linguistic authority of a native speaker and validated against naturally occurring corpus text.

Our central finding is that intra-family Bantu pretraining outperforms massively multilingual XLM-R and French-pivot transfer for Lingala. This suggests a general principle for morphologically rich low-resource African languages. Specifically, typological proximity in the pretraining distribution matters more than pretraining scale. This aligns with the trajectory observed for AfriBERTa and SERENGETI on other African language benchmarks. It also motivates prioritizing language-family-aware model selection over defaulting to the largest available multilingual model.

Future work will expand the dataset with naturally occurring clinical text from community health workers across all three Lingala regional varieties. We will apply parameter-efficient fine-tuning via LoRA for resource-constrained deployment. We will also conduct a prospective evaluation alongside community health workers in the DRC.

Data and Code Availability

All code, the pre-processed dataset, and the Lingala stopword lexicon used in this work are available on GitHub at [2].

Acknowledgments

The authors thank the IndabaX Nigeria organizers, the Nile Basin Scholarship at Alexandria University, and the African Institute for Mathematical Sciences Cameroon Center of Excellence for providing the working environment without which this paper would not be available.

References

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. SERENGETI: Massively multilingual language models for Africa. In *Findings*

2. https://github.com/patricktenga-svg/lingala_triage

of the Association for Computational Linguistics: ACL 2023, pages 1498–1537, Toronto, Canada, 2023.

David Ifeoluwa Adelani et al. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1–25, 2022.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop (NAACL Clinical NLP)*, pages 72–78, Minneapolis, Minnesota, USA, 2019.

Swarthmore College. Ling073: Applied computational linguistics. https://swarthmore.edu/linguistics/ling073. Course materials on low-resource language processing.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8440–8451, Online, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 4171–4186, Minneapolis, Minnesota, 2019.

Ethnologue. Lingala: A language of the DRC. https://www.ethnologue.com/language/lin, 2023. Accessed: 2024-01-15.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, 2021.

Foreign Service Institute. Fsi lingala basic course, 1963. Includes basic grammar, vocabulary, and dialogues for Lingala.

Daniel Hershcovich et al. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6997–7013, Dublin, Ireland, 2022.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, 2014.

Bing Liu. Sentiment analysis and opinion mining. In *Foundations and Trends in Information Retrieval*, volume 2, pages 1–135, 2012.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint*, abs/1907.11692, 2019.

Shamsuddeen Hassan Muhammad et al. AfriSenti: A Twitter sentiment analysis benchmark for african languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Singapore, 2023.

Wilhelmina Nekoto et al. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online, 2020.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. Small data? no problem! pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning (MRL), EMNLP*, pages 12–23, Punta Cana, Dominican Republic, 2021.

David Ogunleye, Bonaventure Masakhane, et al. AfriBERTa: Pretraining for african languages. In *Proceedings of the NeurIPS 2021 Workshop on Resources for African NLP*, 2021.

Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. In *Foundations and Trends in Information Retrieval*, volume 2, pages 1–135, 2008.

Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, and Pascale Fung. Are multilingual models effective in code-switching? In *Proceedings of the 5th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 142–153, Online, 2021.

World Health Organization. Health workforce density by country. https://www.who.int/data/gho/data/themes/topics/health-workforce, 2022. Global Health Observatory. Accessed: 2024-01-15.

David Yocum. Loba Lingala: A learner's guide to spoken Lingala, 2014. Self-published language guide.