N-grams vs. Neural Models for Low-Resource African Languages: When to Switch?

Anonymous ACL submission

Abstract

Practitioners building NLP tools for low-resource African languages face a critical yet understudied decision: should they use simple N-gram-based approaches or invest in neural models given limited training data? While neural language models have largely superseded N-grams for well-resourced languages, their advantage diminishes when training data is scarce. This study presents a systematic empirical comparison across nine typologically diverse African languages spanning four language families (Niger-Congo Bantu, Niger-Congo Volta-Niger, Afro-Asiatic, and an English-based Creole). Using the MasakhaNEWS benchmark, we compare N-gram features with classical classifiers, fine-tuned multilingual pretrained models (XLM-R, AfroXLMR), and small Transformers trained from scratch at controlled training-set sizes from 100 to full available training split. We find three distinct crossover patterns: (1) immediate neural dominance for five languages, (2) persistent N-gram competitiveness for three languages, and (3) complex crossover dynamics for one. Neither language family nor morphological typology predicts the crossover; instead, the target language's cumulative pretraining representation emerges as the dominant factor. We further show that hybrid interpolation of N-gram and neural predictions consistently outperforms both standalone models for eight out of nine models. Character 3-to-5-grams with linear SVM provide a robust N-gram baseline regardless of language typology. Our results offer practical guidelines for model selection in low-resource settings.

1 Introduction

Large-scale pretrained language models such as BERT (Devlin et al., 2019), GPT-3 (Brown et al., 2020), and their multilingual variants (Conneau et al., 2020) have achieved state-of-the-art performance across a broad range of NLP tasks for well-resourced languages. Yet the vast majority of the world's approximately 7,000 languages remain severely underrepresented in both pretraining data and downstream evaluation (Joshi et al., 2020). This disparity is especially pronounced for African languages. More than 2000 languages are spoken across the continent, but collectively they receive a fraction of the research investment directed at English, Chinese, or French (Adebara et al., 2024).

This resource gap poses a concrete, practical question. When building NLP tools with limited training data, should practitioners rely on simple, well-understood N-gram-based models, or invest in more complex neural architectures that have proven dominant in high-resource settings? The answer is not straightforward. Neural models are data-hungry and may underperform simpler baselines when training corpora are small (Pakray et al., 2025). The problem is further compounded for *agglutinative* languages such as those in the Bantu family, in which words are formed by concatenating morphemes, producing explosive vocabulary growth and severe data sparsity even at moderate corpus sizes.

N-gram models have served as the backbone of text classification, language modeling, and information retrieval for decades (Jurafsky and Martin, 2000). Their simplicity, interpretability, and minimal computational requirements make them natural candidates for resource-constrained environments. In contrast, pretrained multilingual models such as XLM-R (Conneau et al., 2020) and Africa-centric adaptations like AfroXLMR (Alabi et al., 2022) offer cross-lingual transfer that can partially compensate for limited target-language data. Baseline experiments on the MasakhaNEWS benchmark (Adelani et al., 2023b) confirm that classical models remain competitive for languages poorly represented in pretraining corpora, yet no prior work has systematically mapped the *crossover point*—the data scale at which neural methods begin to reliably dominate.

We address this gap through a controlled study across **nine** languages spanning four language families. We compare three paradigms for news topic classification: **(i)** N-gram features with classical classifiers, **(ii)** fine-tuned pretrained models (XLM-R and AfroXLMR), and **(iii)** small Transformers trained from scratch, evaluated at training-set sizes of 100, 250, 500, 750, and full available training split.

Contributions. Our contributions are fivefold:

1. We provide a systematic crossover analysis spanning nine African languages across four families, revealing three distinct patterns: immediate neural dominance, persistent N-gram competitiveness, and complex crossover dynamics.

2. We demonstrate that the target language's *cumulative* representation across pretraining stages rather than morphological complexity or language family, is the dominant factor determining the crossover point.

3. We show that small Transformer models trained from scratch consistently underperform both N-gram baselines and fine-tuned models at all scales, confirming that pretraining rather than architecture drives neural advantages in low-resource settings.

4. We demonstrate that hybrid interpolation of N-gram and neural predictions outperforms both standalone models across for eight of nine languages at zero additional training cost.

5. We distill our findings into practical guidelines for practitioners building NLP tools for low-resource languages.

2 Related Work

Low-resource NLP and the resource gap. Joshi et al. (2020) classify roughly 90% of the world's languages as "left-behinds" for which labeled data is scarce or absent. Pakray et al. (2025) identify annotated dataset scarcity as the primary bottleneck for low-resource NLP, noting that transfer learning benefits diminish for languages with minimal pretraining representation. A recurring finding is that classical models paired with N-gram features can match or exceed neural approaches in low-data regimes (Adelani et al., 2023b), motivating our study.

African language NLP. The Masakhane initiative (Adelani et al., 2023a) has catalyzed African NLP, producing MasakhaNEWS (Adelani et al., 2023b), a 16-language news classification benchmark. Baseline experiments show fine-tuned models generally outperform classical classifiers, but the gap narrows for underrepresented languages. Africa-centric pretrained models include AfroXLMR (Alabi et al., 2022), AfriBERTa (Ogueji et al., 2021), Glot500 (Imani et al., 2023), and Cheetah (Adebara et al., 2024).

N-gram versus neural methods. N-gram features, typically TF-IDF vectors, paired with classifiers such as Naive Bayes or SVMs are robust in low-resource settings (Jurafsky and Martin, 2000). On the neural side, pretrained Transformers (Vaswani et al., 2017) offer cross-lingual transfer, though fine-tuning on very small datasets risks overfitting. Adelani et al. (2023b) show that PET (Schick and Schütze, 2021) with 10 examples per label achieves over 90% of fully supervised F1. Despite these advances, no prior work has systematically characterized the crossover across typologically diverse African languages.

Hybrid methods and data augmentation. Interpolating N-gram and neural models has a long history in speech recognition (Jurafsky and Martin, 2000); we investigate whether such hybrids help for low-resource text classification. Recent bootstrapping strategies (Maiti et al., 2025) are complementary to our work, and understanding the crossover point helps practitioners decide whether investing in data expansion is worthwhile.

3 Languages and Dataset

We use the MasakhaNEWS benchmark (Adelani et al., 2023b), the largest news topic classification dataset for African languages. The dataset covers 16 languages with articles sourced primarily from BBC News and labeled across seven categories: business, entertainment, health, politics, religion, sports, and technology. Annotation was performed by native speakers from the Masakhane community, with inter-annotator agreement (Fleiss' κ) ranging from 0.55 to 0.85.

For our experiments, we study nine languages from the benchmark, selected to span four language families and diverse morphological profiles:

Niger-Congo, Bantu (agglutinative): Luganda, Rundi, chiShona, and Kiswahili form words by con-

Table 1: Target languages grouped by family with dataset splits. CC-100 indicates presence in XLM-R's pretraining corpus.

Language	Family	Train	Val	Test	CC-100
Luganda	Bantu	771	110	223	No
Rundi	Bantu	1,117	159	322	No
chiShona	Bantu	1,288	185	369	No
Kiswahili	Bantu	1,658	237	476	Yes
Igbo	Volta-Niger	1,356	194	388	No
Yoruba	Volta-Niger	1,433	205	410	No
Amharic	Semitic	1,311	188	375	Yes
Hausa	Chadic	2,219	317	634	Yes
Oromo	Cushitic	1,128	161	323	Yes
Nig. Pidgin	Creole	1,060	152	303	No

catenating morphemes, producing high vocabulary diversity. Only Kiswahili appears in XLM-R's CC-100 pretraining corpus (∼275M tokens) whereas all four are included in AfroXLMR's continued pretraining.

Niger-Congo, Volta-Niger: Igbo and Yoruba are tonal languages with less agglutinative morphology than Bantu. Neither appears in CC-100 but both are included in AfroXLMR.

Afro-Asiatic: Amharic (Semitic branch, Ge'ez script), Hausa (Chadic, fusional morphology), and Oromo (Cushitic) are all present in CC-100. All three are also included in AfroXLMR's continued pretraining.

Creole: Nigerian Pidgin is an English-based creole with relatively simple morphology. It is absent from CC-100 but included in AfroXLMR.

We construct controlled subsamples of 100, 250, 500, and 750 training examples using stratified sampling to preserve label distributions. We additionally train on the full available training set for each language.

4 Methodology

We compare three modeling paradigms across all nine languages at each data scale (100, 250, 500, 750, and full training set). All experiments are repeated over five random subsampling seeds. Since full-scale experiments involve no subsampling, we use a single seed for full scale training. The validation and test sets are held fixed across all conditions, and we report mean and standard deviation of weighted F1-score on the held-out test set.

4.1 N-gram Baselines

We extract N-gram features using TF-IDF vectorization at both the word and character levels. For word-level features, we use unigrams and bigrams. For character-level features, we use character N-grams of order 3 to 5, which capture sub-word morphological patterns. We pair these features with three classical classifiers:

- **Multinomial Naive Bayes**, a strong probabilistic baseline for text classification with sparse features.
- **Linear SVM** (with ℓ_2 regularization), which performs well in high-dimensional, sparse feature spaces.
- **XGBoost**, a gradient-boosted tree ensemble that can capture non-linear feature interactions.

We tune regularization strength (SVM) and the number of estimators (XGBoost) on the validation set at each data scale. The choice of N-gram range (word vs. character, unigram vs. bigram) is also treated as a hyperparameter.

4.2 Fine-tuned Multilingual Pretrained Models

We fine-tune two pretrained multilingual models:

- **XLM-R$_{\text{base}}$** (Conneau et al., 2020), a 270M-parameter cross-lingual model pretrained on 100 languages using the CC-100 corpus.
- **AfroXLMR$_{\text{base}}$** (Alabi et al., 2022), which adapts XLM-R through continued pretraining on 17 African languages, covering all nine of our target languages (Rundi via the closely related Kinyarwanda).

For each model, we add a classification head (a single linear layer) on top of the `[CLS]` token representation and fine-tune all parameters. We use the AdamW optimizer with a learning rate of 2×10^{-5}, weight decay of 0.01, and linear warmup over the first 10% of steps. We train for a maximum of 20 epochs with early stopping on validation weighted F1 (patience of 10 epochs). Batch sizes are scaled by training-set size (8 at N=100 up to 32 at full scale) to ensure a reasonable number of gradient updates.

4.3 Small Transformers Trained from Scratch

To disentangle the benefit of cross-lingual transfer from the Transformer architecture itself, we train small Transformer models from scratch on the target-language training data alone. We use a 4-layer, 4-head Transformer encoder with a hidden dimension of 256 and a feedforward dimension of 512, amounting to approximately 5M parameters. Tokenization is performed using a SentencePiece unigram model trained on the target-language training data with a vocabulary size of 8,000 and a maximum sequence length of 256 tokens. We use AdamW with a learning rate of 5×10^{-4}, weight decay of 0.01, and train for up to 50 epochs with early stopping on validation weighted F1 (patience of 10 epochs). Batch sizes are scaled by training-set size (16 at $N{=}100$ up to 64 at full scale).

This condition tests whether the Transformer architecture offers benefits over N-gram approaches even without the advantage of large-scale pretraining—a question relevant to practitioners who may not have access to pretrained models covering their target language.

4.4 Hybrid Interpolation

At full training-set scale, we combine predictions from the best N-gram model (SVM) and the best neural model (AfroXLMR):

$$P_{\text{hybrid}}(y \mid x) = \lambda \cdot P_{\text{ngram}}(y \mid x) + (1{-}\lambda) \cdot P_{\text{neural}}(y \mid x) \quad (1)$$

where $\lambda \in [0, 1]$ is tuned on the validation set via grid search over $\{0.0, 0.1, \ldots, 1.0\}$. The hybrid interpolation experiment comes at no additional training cost.

4.5 Evaluation

We evaluate all models using **weighted F1-score**, consistent with the MasakhaNEWS baseline evaluation (Adelani et al., 2023b). Weighted F1 accounts for class imbalance by computing the F1-score for each class and weighting by class frequency.

We consider a neural model to have "crossed over" an N-gram baseline when it achieves a statistically significant improvement (paired t-test, $p < 0.05$) across all five trials.

5 Results

We first summarize the ablation findings that determined our best N-gram configuration, then present the main cross-paradigm comparison.

5.1 N-gram Ablation

Across all nine languages and data scales, **Linear SVM with character 3-to-5-gram TF-IDF features** emerged as the strongest N-gram configuration (full ablation results in Appendix A). This dominance held across all language families, agglutinative Bantu, tonal Volta-Niger, fusional Chadic, and English-based Creole. SVM outperformed both Multinomial Naive Bayes and XGBoost at nearly every language–scale combination. XGBoost consistently underperformed both alternatives, suggesting that gradient-boosted trees offer no advantage over linear classifiers in high-dimensional sparse feature spaces with limited training data.

5.2 Main Comparison Across Paradigms

Table 2 summarizes the core comparison between the best N-gram system (SVM with character 3-to-5-gram features) and the best neural model (AfroXLMR) at two key scales: $N{=}100$ (the most data-constrained setting) and full training set. The complete results across all scales, including XLM-R and from-scratch Transformer baselines, are provided in Appendix A.2.

Table 2: N-gram SVM vs. AfroXLMR at $N{=}100$ (mean over 5 seeds) and full scale. **Bold** = stronger model.

	$N{=}100$		Full	
Language	SVM	AfroXLMR	SVM	AfroXLMR
Luganda	**0.769**	0.546	0.877	**0.890**
Rundi	0.746	**0.861**	**0.897**	0.895
chiShona	**0.895**	0.895	**0.935**	0.932
Kiswahili	0.704	**0.783**	0.825	**0.847**
Igbo	0.694	**0.763**	0.829	**0.853**
Yoruba	0.797	**0.828**	**0.900**	0.895
Amharic	0.854	**0.895**	**0.921**	0.920
Hausa	0.807	0.824	0.879	**0.884**
Oromo	**0.776**	0.750	**0.897**	0.865
Nig. Pidgin	0.892	**0.946**	0.967	**0.978**

At $N{=}100$, AfroXLMR leads for six languages, N-gram SVM leads for two (Luganda and Oromo), and the two are tied for chiShona. However, at full scale the picture shifts: N-gram SVM leads or matches AfroXLMR for five languages (Rundi, chiShona, Yoruba, Amharic, Oromo), demonstrating that with sufficient task-specific data, simple models can close the gap opened by pretraining. The full per-scale comparison (Appendix A.2) reveals the trajectory of this convergence for each language.

Table 3: Crossover point by language, determined by paired t-test ($p < 0.05$) across 5 seeds.

Language	Family	Crossover N	Key finding
Neural-dominant (5 languages)			
Rundi	Bantu	100	$p<.001$ all scales
Kiswahili	Bantu	100	$p<.05$ all scales
Igbo	V-Niger	100	$p<.04$ all scales
Nig. Pid.	Creole	100	$p<.03$ all scales
Amharic	Semitic	100*	$p=.021$; not sig. 250–750
N-gram competitive (3 languages)			
chiShona	Bantu	None	SVM sig. at 500 ($p=.022$)
Hausa	Chadic	None	No sig. diff. any scale
Oromo	Cushitic	None	SVM sig. at 750 ($p=.006$)
Complex (2 languages)			
Luganda	Bantu	750	SVM sig. at 100; late cross.
Yoruba	V-Niger	—	Reversal: see text

5.3 Crossover Analysis

Table 3 presents the crossover results determined by paired t-tests. Three distinct patterns emerge:

Neural-dominant. Five languages show immediate neural dominance. For Rundi, Kiswahili, Igbo, and Nigerian Pidgin, AfroXLMR is significantly better at all seeded scales. Amharic shows significance at N=100 ($p = 0.021$) but the gap narrows at larger scales, suggesting N-grams catch up as data grows.

N-gram competitive. For chiShona, Hausa, and Oromo, N-gram SVM matches or exceeds AfroXLMR at all scales. chiShona's SVM is *significantly better* at N=500 ($p = 0.022$). Oromo's SVM is significantly better at N=750 ($p = 0.006$). For Hausa, the two approaches show no significant difference at any scale.

Complex crossover. Luganda shows SVM significantly leading at N=100 ($p = 0.019$), with AfroXLMR's first significant advantage not until N=750 ($p = 0.049$). Yoruba exhibits a unique *reversal*: AfroXLMR significantly wins at N=100 ($p = 0.004$), SVM significantly wins back at N=250 ($p = 0.018$), and neither dominates at 500–750.

Language family does not predict the crossover. Bantu languages span all three patterns. Volta-Niger languages show opposite patterns (Igbo: neural; Yoruba: reversal). Afro-Asiatic languages are similarly split.

5.4 The Role of Africa-Specific Pretraining

The crossover aligns with pretraining representation, though the relationship is nuanced. Four of our nine languages appear in XLM-R's CC-100 corpus (Kiswahili, Amharic, Hausa, Oromo), yet their crossover patterns differ dramatically: Kiswahili and Amharic show neural dominance, while Hausa and Oromo remain N-gram competitive. This suggests that mere presence in CC-100 is insufficient signifying the *amount* and *quality* of pretraining data matters.

AfroXLMR covers all nine languages but with vastly different data volumes, from 67K sentences (Luganda) to 2.7M (chiShona). The relationship between AfroXLMR data volume and crossover is not monotonic. chiShona received the *most* AfroXLMR data yet N-grams dominate, suggesting that character N-grams capture chiShona's regular morphological patterns so effectively that even extensive pretraining cannot provide an edge. Conversely, Rundi received only 304K sentences but benefits from strong cross-lingual transfer via the closely related Kinyarwanda.

For Kiswahili, present in *both* CC-100 ($\sim$275M tokens) and AfroXLMR, neural models dominate from N=100. For Luganda, absent from CC-100 with only 67K AfroXLMR sentences, N-grams lead until N=750. The cumulative effect of both pretraining stages, rather than either alone, appears to determine whether neural models can gain a decisive advantage.

5.5 From-Scratch Transformers

Small Transformer models trained from scratch (5M parameters, 4-layer encoder) **consistently underperform both N-gram baselines and fine-tuned pretrained models at every scale for every language**.

This finding confirms that the advantage of Transformer-based models in low-resource settings derives primarily from *pretraining*—the cross-lingual knowledge transferred from high-resource languages—rather than from the Transformer architecture itself. For practitioners without access to pretrained models covering their target language, N-gram baselines with character-level features are strictly preferable to training small Transformers from scratch.

Table 4: Hybrid interpolation at full scale. λ tuned on validation set. Δ: improvement over stronger standalone model.

Language	λ	Hybrid	Best solo	Δ
Luganda	0.8	**0.905**	0.900	+0.005
Rundi	0.8	**0.910**	0.907	+0.004
chiShona	0.6	**0.946**	0.935	+0.011
Kiswahili	0.7	**0.860**	0.850	+0.010
Igbo	0.7	**0.865**	0.851	+0.014
Yoruba	0.8	**0.920**	0.903	+0.017
Oromo	0.9	**0.917**	0.914	+0.003
Nig. Pidgin	0.5	**0.984**	0.981	+0.003
Hausa	0.8	**0.891**	0.890	+0.001
Amharic	0.0	0.923	0.925	−0.002

5.6 Hybrid Interpolation

Table 4 presents the hybrid results at full scale. The hybrid outperforms both standalone models for eight of nine languages; for Amharic, the val-tuned $\lambda = 0.0$ (pure neural) yields no improvement.

The optimal λ reflects the underlying crossover pattern: N-gram-competitive languages favour high λ (Oromo: 0.9), neural-dominant languages favour lower values (Nigerian Pidgin: 0.5). The highest F1 in our entire study is the Nigerian Pidgin hybrid at 0.984.

6 Discussion

6.1 Practical Guidelines

Our results yield the following decision framework:

1. **Check Africa-specific pretrained model coverage.** If AfroXLMR covers the target language with strong pretraining data, fine-tune even with 100 examples.

2. **Poor or uncertain coverage? Start with N-gram SVM.** Character N-gram features with linear SVM match or outperform neural models for three of nine languages at all scales tested.

3. **Use character 3-to-5-grams as the default.** This configuration dominates across all language families, including isolating, fusional, and creole languages—not just agglutinative ones.

4. **Apply hybrid interpolation.** A weighted average of N-gram and neural predictions ($\lambda = 0.5$–0.9) consistently improves over either model alone at zero cost.

5. **Do not train Transformers from scratch** with fewer than 2,000 samples—pretraining, not architecture, drives neural advantages.

7 Conclusion

We presented a systematic comparison of N-gram and neural approaches for news topic classification across nine typologically diverse African languages at controlled training-set sizes. Our central finding is that **there is no universal crossover point**: the answer ranges from immediate neural dominance (Rundi, Kiswahili, Igbo, Nigerian Pidgin) to persistent N-gram competitiveness (chiShona, Hausa, Oromo) to complex dynamics (Luganda, Yoruba). Neither language family nor morphological typology predicts the crossover; instead, *cumulative pretraining representation* emerges as the dominant factor. Hybrid interpolation consistently outperforms both standalone models, and character 3-to-5-grams with linear SVM provide a robust baseline regardless of typology. These findings offer practical model-selection guidelines for the growing community building NLP tools for African languages.

Limitations

Our study has several limitations that constrain the generalizability of our findings. First, we evaluate on a single task (news topic classification) using a single benchmark (MasakhaNEWS). The crossover dynamics may differ for other tasks such as named entity recognition, sentiment analysis, or machine translation. Second, the hybrid interpolation experiment was conducted at full scale with a single seed; multi-seed evaluation remains future work. Third, our from-scratch Transformer uses a single architecture; different configurations might yield different results. Fourth, the training-set sizes in MasakhaNEWS are small even at full scale, preventing study of the crossover at larger data scales. Finally, while we span four language families, the sample within each family is small (e.g., two Volta-Niger languages), limiting within-family generalization.

Ethical Considerations

This research uses the publicly available MasakhaNEWS benchmark dataset, which was collected and annotated by the Masakhane community with informed consent from annotators. No personally identifiable information is contained

in the data. Our work aims to provide practical guidance for NLP practitioners working on low-resource African languages, contributing to more equitable access to language technology. We acknowledge that our findings are specific to the nine languages studied and should not be overgeneralized to all African or low-resource languages without further investigation. We used AI-assisted tools (Claude Code) for coding support during experiment implementation. All code was reviewed and validated by the authors.

References

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Guilherme Alcoba Inciarte. 2024. Cheetah: Natural language generation for 517 African languages. *arXiv preprint arXiv:2401.01813*.

David Ifeoluwa Adelani, Tunde Ajayi, Ife Adebara, Abraham Ayeni, Happy Buzaaba, Cheikh M. Bamba Dione, and 1 others. 2023a. Masakhane—machine translation for Africa. *Transactions of the Association for Computational Linguistics*, 11:738–755.

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure F. P. Dossou, Akintunde Oladipo, Doreen Nixdorf, Chris Chinenye Emezue, and 1 others. 2023b. MasakhaNEWS: News topic classification for African languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 144–159, Nusa Dua, Bali. Association for Computational Linguistics.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. 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. International Committee on Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ayyoob Imani, Silvia Severini, Masoud Jalili Sabet, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language technology to 500 languages and beyond. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 592–616, Toronto, Canada. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Buber, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Daniel Jurafsky and James H. Martin. 2000. *Speech and Language Processing*, 2nd edition. Prentice Hall.

Agniva Maiti, Manya Pandey, and Murari Mandal. 2025. NagaNLP: Bootstrapping NLP for low-resource Nagamese Creole with human-in-the-loop synthetic data. *Preprint*, arXiv:2512.12537.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. 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. Association for Computational Linguistics.

Partha Pakray, Alexander Gelbukh, and Sivaji Bandyopadhyay. 2025. Natural language processing applications for low-resource languages. *Natural Language Processing*, 31(2):183–197.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

A N-gram Ablation Details

A.1 Classifier Comparison

Table 5 reports per-classifier results for the N-gram track. Linear SVM with ℓ_2 regularization is the strongest classifier overall, dominating Naive Bayes and XGBoost at nearly every language–scale combination.

A.2 Main Comparison Across Paradigms

Table 6 presents the central results of our study: weighted F1 scores for the best N-gram system, both fine-tuned pretrained models, and the from-scratch Transformer across all languages and training-set sizes.

A.3 Feature Type Selection

Analysis of the best-performing feature type across all N-gram trials across nine languages reveals a strong preference for character-level features. Character 3-to-5-grams (`char_3to5`) were selected as the best feature type in the majority of trials, particularly for SVM and XGBoost classifiers. Word unigrams were occasionally competitive at N=100 for some languages (notably Luganda and Rundi), where the very small training set may not provide sufficient character-level coverage. Word bigrams appeared as the best feature for a minority of Rundi and Kiswahili trials. The dominance of character N-grams confirms that sub-word features are effective across typologically diverse languages, not just agglutinative ones. This robustness held for isolating (Yoruba, Igbo), fusional (Hausa), and creole (Nigerian Pidgin) languages as well as agglutinative Bantu languages

B Fine-tuning Detailed Results

Table 7 presents the full per-model results for fine-tuned pretrained models. The gap between AfroXLMR and XLM-R varies substantially across languages and scales, reflecting differences in pretraining coverage.

Several patterns are notable. AfroXLMR exhibits high variance for Luganda at N=100 (std = 0.129), indicating that fine-tuning is unstable with very few examples when pretraining coverage is poor. By contrast, AfroXLMR on Rundi at N=100 achieves 0.861 ± 0.011—both high and stable—suggesting that Rundi benefits more from AfroXLMR's pretraining. For Kiswahili, XLM-R occasionally matches or surpasses AfroXLMR (e.g., 0.786 vs. 0.783 at N=100; 0.851 vs. 0.834 at full scale), indicating that when a language is well-represented in XLM-R's pretraining data, Africa-specific adaptation provides diminishing returns.

The AfroXLMR advantage over XLM-R is largest for Luganda (+0.181 F1 at N=100, +0.321 at N=250) and decreases monotonically as training data grows, consistent with the intuition that cross-lingual transfer is most valuable when task-specific data is scarce. Oromo exhibits extremely high XLM-R variance at N=100 (std = 0.152), despite being present in CC-100, suggesting that its CC-100 representation may be too small for reliable transfer. Nigerian Pidgin shows the smallest AfroXLMR–XLM-R gap across all languages, consistent with its English-based vocabulary enabling strong transfer from XLM-R's substantial English pretraining data.

C From-Scratch Transformer Detailed Results

Table 8 presents the full results for the 4-layer Transformer encoder trained from random initialization.

The from-scratch Transformer underperforms both N-gram SVM and fine-tuned pretrained models at every language–scale condition. The performance gap is most severe for Kiswahili (0.514 at N=100 vs. 0.704 for N-gram SVM and 0.786 for XLM-R), suggesting that without pretraining, the Transformer architecture struggles most on the language with the lowest type-to-token ratio, where its limited vocabulary of 8,000 SentencePiece tokens may be particularly ill-suited.

The variance is also higher for from-scratch models than for fine-tuned models at small scales (e.g., Luganda std = 0.051 at N=100), reflecting the instability of training a 5M-parameter model on very small datasets without the regularizing effect of pretrained weights.

D Training Hyperparameters

Table 9 summarizes the hyperparameter configurations used across all three experimental tracks.

Table 9: Hyperparameter settings for all experimental tracks.

Parameter	Fine-tuning	Scratch
Architecture	XLM-R / AfroXLMR (270M)	4L, 4H encoder (5M)
Tokenizer	SP 250K vocab	Per-lang SP 8K
Max seq. len	256	256
Optimizer	AdamW	AdamW
Learning rate	2×10^{-5}	5×10^{-4}
Weight decay	0.01	0.01
LR schedule	Linear 10% warmup	Linear 10% warmup
Max epochs	20	50
Early stop	Pat. 10 (val w-F1)	Pat. 10 (val w-F1)
Hidden / FF	768 / 3072	256 / 512
Attn. heads	12	4
Dropout	0.1	0.1
Eval batch	64	128
Train batch	8/16/16/32/32	16/32/32/64/64
Deterministic	`cudnn_det.`	`cudnn_det.`
N-gram baselines		
Features	word uni/bi, char 3-5, 2-6, combined	
NB α	$\{0.01, 0.1, 0.5, 1.0\}$	
SVM C	$\{0.1, 1.0, 10.0, 100.0\}$	
XGB trees	$\{100, 200, 300, 500\}$	

Computing Infrastructure. N-gram experiments were run on Modal cloud compute. Fine-tuning and from-scratch Transformer experiments were conducted on Google Colab Pro using a single NVIDIA A100 GPU.

Table 5: N-gram weighted test F1 by classifier (mean ± std over 5 seeds; single seed at full scale). Best classifier per condition in **bold**.

Language	N	Naive Bayes	SVM	XGBoost
Luganda	100	0.772 ± .028	**0.769** ± .033	0.680 ± .020
	250	0.815 ± .016	**0.852** ± .011	0.787 ± .008
	500	0.854 ± .009	**0.871** ± .013	0.825 ± .011
	750	0.865 ± .004	**0.880** ± .004	0.847 ± .011
	full	0.865	**0.877**	0.839
Rundi	100	0.721 ± .026	**0.746** ± .025	0.700 ± .022
	250	0.770 ± .016	**0.815** ± .005	0.746 ± .030
	500	0.810 ± .026	**0.848** ± .012	0.794 ± .023
	750	0.835 ± .009	**0.853** ± .011	0.812 ± .025
	full	0.857	**0.897**	0.861
chiShona	100	0.886 ± .021	**0.895** ± .013	0.830 ± .016
	250	0.906 ± .007	**0.913** ± .007	0.880 ± .009
	500	0.916 ± .009	**0.929** ± .006	0.899 ± .012
	750	0.917 ± .014	**0.924** ± .011	0.912 ± .005
	full	**0.935**	0.916	0.913
Kiswahili	100	0.656 ± .035	**0.704** ± .021	0.618 ± .022
	250	0.705 ± .015	**0.767** ± .019	0.675 ± .024
	500	0.755 ± .014	**0.797** ± .013	0.727 ± .017
	750	0.768 ± .005	**0.804** ± .011	0.758 ± .012
	full	0.790	**0.825**	0.785
Igbo	100	0.656 ± .049	**0.694** ± .032	0.584 ± .023
	250	0.728 ± .025	**0.773** ± .017	0.674 ± .037
	500	0.783 ± .011	**0.799** ± .018	0.749 ± .028
	750	0.797 ± .006	**0.816** ± .004	0.763 ± .029
	full	0.804	**0.829**	0.768
Yoruba	100	0.766 ± .019	**0.797** ± .007	0.748 ± .025
	250	0.832 ± .018	**0.868** ± .005	0.813 ± .015
	500	0.856 ± .017	**0.879** ± .012	0.832 ± .009
	750	0.859 ± .021	**0.893** ± .006	0.855 ± .015
	full	0.878	**0.900**	0.866
Amharic	100	**0.854** ± .008	**0.854** ± .007	0.758 ± .045
	250	0.884 ± .007	**0.892** ± .009	0.836 ± .016
	500	0.900 ± .005	**0.901** ± .012	0.851 ± .014
	750	0.902 ± .001	**0.910** ± .009	0.862 ± .017
	full	0.904	**0.921**	0.868
Hausa	100	0.764 ± .026	**0.807** ± .015	0.649 ± .020
	250	0.812 ± .006	**0.843** ± .009	0.743 ± .017
	500	0.837 ± .013	**0.865** ± .008	0.801 ± .005
	750	0.849 ± .011	**0.872** ± .012	0.818 ± .010
	full	0.873	**0.879**	0.865
Oromo	100	0.738 ± .028	**0.776** ± .020	0.707 ± .013
	250	0.779 ± .009	**0.832** ± .016	0.758 ± .024
	500	0.838 ± .008	**0.862** ± .007	0.815 ± .018
	750	0.856 ± .009	**0.884** ± .009	0.839 ± .014
	full	0.870	**0.897**	0.857
Nig. Pidgin	100	0.879 ± .021	**0.892** ± .016	0.814 ± .011
	250	0.909 ± .009	**0.929** ± .007	0.874 ± .017
	500	0.928 ± .009	**0.954** ± .007	0.904 ± .013
	750	0.934 ± .008	**0.960** ± .010	0.931 ± .003
	full	0.922	**0.967**	0.943

Table 6: Weighted test F1 across paradigms by language and training-set size. N-gram reports the best classifier per seed averaged over 5 seeds; neural results are mean ± std over 5 seeds (single seed at full scale). **Bold** = best method at each condition.

Language	N	N-gram	AfroXLMR	XLM-R	Scratch
Luganda	100	**0.785**	0.546 ± .129	0.365 ± .049	0.635 ± .051
	250	**0.852**	0.798 ± .052	0.477 ± .056	0.728 ± .016
	500	0.870	**0.882** ± .012	0.649 ± .047	0.798 ± .019
	750	0.880	**0.887** ± .005	0.692 ± .047	0.803 ± .027
	full	**0.877**	0.859	0.727	0.822
Rundi	100	0.749	**0.861** ± .011	0.629 ± .062	0.627 ± .051
	250	0.815	**0.896** ± .008	0.742 ± .033	0.720 ± .023
	500	0.848	**0.895** ± .009	0.803 ± .006	0.758 ± .031
	750	0.853	**0.891** ± .009	0.822 ± .020	0.793 ± .018
	full	**0.897**	0.892	0.834	0.826
chiShona	100	**0.896**	0.895 ± .017	0.759 ± .025	0.759 ± .015
	250	0.913	**0.918** ± .009	0.833 ± .017	0.853 ± .018
	500	**0.929**	0.922 ± .007	0.866 ± .009	0.863 ± .020
	750	0.924	**0.928** ± .004	0.874 ± .010	0.869 ± .008
	full	**0.935**	0.932	0.905	0.873
Kiswahili	100	0.704	0.783 ± .054	**0.786** ± .020	0.514 ± .037
	250	0.767	**0.833** ± .015	0.816 ± .014	0.593 ± .036
	500	0.797	**0.836** ± .010	0.831 ± .015	0.666 ± .021
	750	0.804	**0.843** ± .007	0.832 ± .014	0.707 ± .010
	full	0.825	0.834	**0.851**	0.744
Igbo	100	0.694	**0.763** ± .021	0.615 ± .022	0.551 ± .016
	250	0.773	**0.811** ± .023	0.653 ± .030	0.663 ± .034
	500	0.799	**0.829** ± .017	0.737 ± .023	0.713 ± .007
	750	0.816	**0.848** ± .007	0.746 ± .018	0.739 ± .023
	full	0.829	**0.853**	0.778	0.789
Yoruba	100	**0.797**	0.828 ± .007	0.582 ± .056	0.679 ± .043
	250	**0.868**	0.860 ± .007	0.723 ± .039	0.749 ± .017
	500	**0.879**	0.877 ± .006	0.781 ± .009	0.786 ± .039
	750	**0.893**	0.889 ± .005	0.789 ± .010	0.825 ± .030
	full	**0.900**	0.895	0.827	0.853
Amharic	100	0.854	**0.895** ± .022	0.869 ± .022	0.703 ± .037
	250	0.892	**0.897** ± .017	0.881 ± .010	0.804 ± .026
	500	0.901	**0.913** ± .006	0.884 ± .009	0.834 ± .030
	750	0.910	**0.923** ± .004	0.892 ± .013	0.850 ± .012
	full	**0.921**	0.920	0.900	0.890
Hausa	100	**0.807**	0.823 ± .051	0.771 ± .020	0.564 ± .048
	250	0.843	**0.852** ± .013	0.810 ± .015	0.678 ± .023
	500	**0.865**	0.863 ± .004	0.842 ± .009	0.750 ± .027
	750	0.872	**0.876** ± .009	0.852 ± .012	0.763 ± .024
	full	0.879	**0.884**	0.871	0.821
Oromo	100	**0.776**	0.750 ± .031	0.473 ± .152	0.648 ± .029
	250	**0.832**	0.831 ± .027	0.736 ± .045	0.765 ± .011
	500	**0.862**	0.852 ± .019	0.815 ± .012	0.769 ± .017
	750	**0.884**	0.865 ± .008	0.815 ± .014	0.794 ± .024
	full	**0.897**	0.865	0.837	0.829
Nig. Pidgin	100	0.892	**0.946** ± .012	0.937 ± .018	0.764 ± .027
	250	0.929	**0.964** ± .006	0.955 ± .010	0.870 ± .019
	500	0.954	**0.971** ± .004	0.966 ± .003	0.890 ± .015
	750	0.960	**0.972** ± .005	0.970 ± .003	0.900 ± .019
	full	0.967	**0.978**	0.970	0.921

Table 7: Fine-tuned pretrained model test weighted F1 by language and training-set size (mean ± std over 5 seeds; single seed at full scale).

Model	Language	100	250	500	750	full
AfroXLMR	Luganda	0.546 ± .129	0.798 ± .052	0.882 ± .012	0.887 ± .005	0.859
	Rundi	0.861 ± .011	0.896 ± .008	0.895 ± .009	0.891 ± .009	0.892
	chiShona	0.895 ± .017	0.918 ± .009	0.922 ± .007	0.928 ± .004	0.932
	Kiswahili	0.783 ± .054	0.833 ± .015	0.836 ± .010	0.843 ± .007	0.834
XLM-R	Luganda	0.365 ± .049	0.477 ± .056	0.649 ± .047	0.692 ± .047	0.727
	Rundi	0.629 ± .062	0.742 ± .033	0.803 ± .006	0.822 ± .020	0.834
	chiShona	0.759 ± .025	0.833 ± .017	0.866 ± .009	0.874 ± .010	0.905
	Kiswahili	0.786 ± .020	0.816 ± .014	0.831 ± .015	0.832 ± .014	0.851
AfroXLMR	Igbo	0.763 ± .021	0.811 ± .023	0.829 ± .017	0.848 ± .007	0.853
	Yoruba	0.828 ± .007	0.860 ± .007	0.877 ± .006	0.889 ± .005	0.895
	Amharic	0.895 ± .022	0.897 ± .017	0.913 ± .006	0.923 ± .004	0.920
	Hausa	0.823 ± .051	0.852 ± .013	0.863 ± .004	0.876 ± .009	0.884
	Oromo	0.750 ± .031	0.831 ± .027	0.852 ± .019	0.865 ± .008	0.865
	Nig. Pidgin	0.946 ± .012	0.964 ± .006	0.971 ± .004	0.972 ± .005	0.978
XLM-R	Igbo	0.615 ± .022	0.653 ± .030	0.737 ± .023	0.746 ± .018	0.778
	Yoruba	0.582 ± .056	0.723 ± .039	0.781 ± .009	0.789 ± .010	0.827
	Amharic	0.869 ± .022	0.881 ± .010	0.884 ± .009	0.892 ± .013	0.900
	Hausa	0.771 ± .020	0.810 ± .015	0.842 ± .009	0.852 ± .012	0.871
	Oromo	0.473 ± .152	0.736 ± .045	0.815 ± .012	0.815 ± .014	0.837
	Nig. Pidgin	0.937 ± .018	0.955 ± .010	0.966 ± .003	0.970 ± .003	0.970

Table 8: From-scratch Transformer test weighted F1 (mean ± std over 5 seeds at subsampled scales; mean ± std over 5 seeds at full scale).

Language	100	250	500	750	full
Luganda	0.370 ± .095	0.606 ± .131	0.653 ± .085	0.693 ± .067	0.762 ± .125
Rundi	0.424 ± .213	0.612 ± .101	0.704 ± .044	0.696 ± .065	0.786 ± .078
chiShona	0.452 ± .165	0.748 ± .148	0.750 ± .130	0.792 ± .037	0.877 ± .018
Kiswahili	0.251 ± .174	0.466 ± .095	0.457 ± .141	0.556 ± .042	0.727 ± .032
Igbo	0.551 ± .016	0.663 ± .034	0.713 ± .007	0.739 ± .023	0.789 ± .015
Yoruba	0.679 ± .043	0.749 ± .017	0.786 ± .039	0.825 ± .030	0.853 ± .012
Amharic	0.703 ± .037	0.804 ± .026	0.834 ± .030	0.850 ± .012	0.890 ± .008
Hausa	0.564 ± .048	0.678 ± .023	0.750 ± .027	0.763 ± .024	0.821 ± .015
Oromo	0.648 ± .029	0.765 ± .011	0.769 ± .017	0.794 ± .024	0.829 ± .007
Nig. Pidgin	0.764 ± .027	0.870 ± .019	0.890 ± .015	0.900 ± .019	0.921 ± .019