Automatic Correction of Writing Anomalies in Hausa Texts

Ahmad Mustapha Wali and Sergiu Nisioi*
Human Language Technologies Research Center
Faculty of Mathematics and Computer Science
University of Bucharest
ahmadmwali@gmail.com
sergiu.nisioi@unibuc.ro

Abstract

Hausa texts are often characterized by writing anomalies such as incorrect character substitutions and spacing errors, which sometimes hinder natural language processing (NLP) applications. This paper presents an approach to automatically correct the anomalies by fine-tuning transformer-based models. Using a corpus gathered from several public sources, we create a large-scale parallel dataset of over 400,000 noisy-clean Hausa sentence pairs by introducing synthetically generated noise to mimic realistic writing errors. Moreover, we finetune several multilingual and African language models, including M2M100, AfriTeVA, NCAIR1/N-ATLaS, UBC-NLP/cheetah-base, and other variants of BART and T5 for this correction task. Our experimental results demonstrate that models such as M2M100 achieve state-of-the-art results despite their smaller size and distinct pretraining, and that correcting errors can have a significant impact in improving downstream tasks such as text classification, machine translation, question answering, and LLM prompting in general. This research provides a methodology, a publicly available dataset, and a comparison of models to improve Hausa text quality, thereby advancing NLP capabilities for the language and offering transferable insights for other low-resource languages.

1 Introduction

Hausa (ha, ISO 639-1) belongs to the Chadic branch of the Afro-Asiatic language family and is widely spoken in West Africa. The language has deep historical roots and it is connected to other major Afro-Asiatic languages (from the Semitic, Cushitic, and Berber families) that span regions of Africa and West Asia (Newman, 2000). As a major regional language, Hausa has a large speaker base of approximately 80 million people, primarily residing in West Africa (Eberhard et al., 2023). The majority of Hausa speakers in Nigeria and Niger are first-language users (Eberhard et al., 2023). Beyond these areas, Hausa functions as a lingua franca in regions where it is not the native language (Eberhard et al., 2023; Blench, 2012), being a key language for inter-group communication and regional integration across Nigeria's Middle Belt, northern Ghana, and the Benin Republic.

The increased use of Hausa in digital communication, particularly on social media, has raised new concerns about orthographic variations and the use of informal languages (Zakari et al., 2021). The prevalence of writing errors in digitally available texts further exacerbates the lack of high-quality data for Hausa NLP. These irregularities can be characterized as either character substitution, where certain Hausa characters (ɓ, ɗ, ƙ, and ƴ) are replaced with Standard English consonants, or spacing issues, which involve both the removal and addition of spaces between words. Although these writing anomalies are easily disambiguated by humans, they pose significant challenges to NLP models.

For example, the sentence "abincin ba shi da daɗi" (Eng. *the food is not delicious*) is commonly written as "abincin bashi da dadi," (Eng. *food bought on credit is delicious*) a shift in semantics that completely affects the meaning of the sentence. Character swaps, such as "Wannan ƴa ta ce" (Eng. *this is my daughter*) becoming "Wannan ya ta ce" (Eng. *this is my elder sister*), introduce confusion and noise into the data. Although model architecture and training methods are important, the performance of NLP systems also depends on the quality of the training data (Hedderich et al., 2021). This factor is especially relevant for low-resource languages such as Hausa, where data availability is limited.

Our work fills several gaps for Hausa NLP, firstly by building a large synthetic dataset of writing anomalies using a small seed of social-media data, secondly by building the first set of tools for cor-

*Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9514–9528

recting writing anomalies in Hausa texts based on transformer architectures (Hedderich et al., 2021; Shode et al., 2023), and thirdly by evaluating the impact of correcting such errors in downstream tasks. To summarize our contributions, we provide:

- An analysis and categorization of common writing anomalies prevalent in digital Hausa texts.
- A new dataset comprising of over ₦00,000 samples, constructed by synthetically generating noise, mimicking common Hausa writing errors, thereby creating noisy-clean parallel pairs[1] suitable for text correction tasks and broader Hausa NLP research.
- We train and evaluate several Transformer-based models for the automatic detection and correction of these identified Hausa writing anomalies.
- We present a quantitative evaluation of text quality improvements achieved by our correction models, measured using standard text evaluation metrics and assessed through their impact on downstream tasks.

2 Related Work

Challenges in Low-Resource Language Processing

Data scarcity is a major challenge in low-resource NLP, and low-resource languages have a distinct disadvantage because they typically lack the digitized text, annotated corpora, and parallel data required for machine translation (Lusito et al., 2023). This lack of training data directly decreases the ability of an NLP model to learn linguistic patterns and generalize efficiently, lowering overall performance (Hedderich et al., 2021; Haddow et al., 2022). In addition, reduced digital literacy rates, as well as the strong oral traditions associated with certain low-resource languages, further limit corpus generation and the representativeness of such data (Joshi et al., 2020; Adelani et al., 2025a).

Numerous low-resource languages have intricate morphology and complex lexicon systems, which pose further challenges (Wiemerslage et al., 2022; Oncevay et al., 2022). In particular, Hausa is an agglutinative language, creating challenges less common in high-resource contexts such as English or Romance languages (Nirenburg, 2009; Uwaezuoke and Anachunam, 2023).

Furthermore, low-resource languages have limited accessible tools and resources, unlike high-resource languages, which have an advantage from extensive ecosystems of pretrained language models, comprehensive NLP libraries, and easily accessible infrastructure, including standard keyboards and encoding techniques. Materials in low-resource languages often require researchers to allocate time and funds judiciously to establish tools and resources from scratch (Joshi et al., 2020).

Orthographic variations and informal language use, especially in digital texts from low-resource environments (Nekoto et al., 2020), further exacerbate these issues. Consequently, these diverse challenges directly impact the effectiveness of NLP tasks for low-resource languages. For machine translation, for example, the lack of parallel corpora leads to a lower translation quality compared to pairs of high-resource languages (Imankulova et al., 2019). Similarly, a lack of annotated data and natural language complexity restrict operations such as part-of-speech tagging, parsing, and named entity recognition, decreasing precision and robustness (Imankulova et al., 2019; Karakanta et al., 2017).

Hausa Natural Language Processing: Status and Specific Challenges

NLP tools for Hausa have lately shown considerable growth (Bashir et al., 2017; Abubakar et al., 2019; Akinfaderin, 2020; Rakhmanov and Schlippe, 2022; Parida et al., 2023; Ahmad et al., 2024; Alabi et al., 2025; Sani et al., 2025; Uemura et al., 2026) owing in part to the recent series of AfricaNLP Workshops (Lignos et al., 2025; Chimoto et al., 2026) and the sustained effort of the Masakhane community (Adelani et al., 2021; Dione et al., 2023; Muhammad et al., 2025). Despite the recent growth of resources, challenges still exist. Hausa's morphological system provides specific computational problems by combining concatenative and non-concatenative processes (Crysmann, 2016). This intricacy is notable in its verbal system, where derivation from a single root can create several forms that need robust morphological analytic methods. For example, the root word "karya" (*to break*) has several forms, including karyawa (*breaking*), karye (*broken*), kakkarye (*broken several times*), kari (*breaking in noun form*), karyayye (*a broken object(m)*) and so on. Hausa's interaction of tone and meaning adds yet another level of complexity that existing

[1] Released under CC-BY-4.0 license.

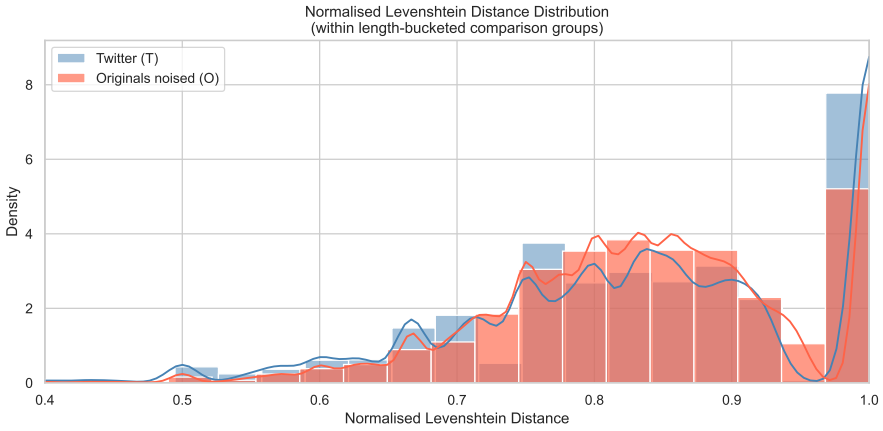

Figure 1: Our dataset's characteristics are calibrated so that its normalized Levenshtein distance distribution resembles that of the naturally-noisy Twitter dataset. This process results in a parallel corpus of noisy-to-clean sentence pairs. Normalized Levenshtein Distance Distribution between the twitter dataset and the synthetically-generated noisy dataset calibrate with the parameters from Table 2.

NLP systems are sometimes unable to comprehend (Crysmann, 2016; Zakari et al., 2021), such as how the word "baba" can mean any of father, mother, or pal depending on the intonation used.

Transformer-based models have opened new pathways for Hausa NLP, with experiments using multilingual African models (Belay et al., 2025) demonstrating potential applications in named entity recognition and sentiment analysis, among others (Muhammad et al., 2022; Shode et al., 2023; Rakhmanov and Schlippe, 2022). However, owing to inadequate representation in pretraining data, their performance in Hausa consistently lags behind that of high-resource languages (Zakari et al., 2021). Limitations in computational resources and the lack of high-quality training data (Salahudeen et al., 2023) exacerbate the challenges of fine-tuning these models.

These technical challenges are amplified by broader infrastructural and institutional constraints. While many Nigerian and Nigerien institutions provide computer science and linguistics courses, specialized education in computational Hausa language processing is still rare (Zakari et al., 2021). Some of these issues are starting to be addressed by recently community-led, multinational cooperative projects. HausaNLP (Emezue et al., 2023; Ahmad et al., 2025; Hussen et al., 2025) provides a platform for coordinating research, developing models and datasets, and sharing resources among institutions focused on Hausa and other African languages.

3 Data Collection and Preparation

The foundation of our work is a large Hausa corpus aggregated from multiple publicly available datasets. The specific datasets incorporated include Hausa Wikipedia (a crawl of the existing data), Wikimedia Hausa, and a collection of other miscellaneous Hausa texts. Furthermore, several texts of different genres from the MultiplEYE project (Kasperė et al., 2026; Nisioi et al., 2026) have been translated into Hausa following the official project guidelines (Hollenstein et al., 2026). These texts are included as additional out-of-domain (**OOD**) data, since they have not been previously published online and can be used for a fair comparison of models. The texts are multi-genre, and cover two argumentative texts from Programme for International Student Assessment (PISA), two institutional texts (Declaration of Human Rights and a text from the European Commission), seven literary texts, and two pop-science texts. In total, we include 10,000 words of high-quality translated texts into Hausa.

After merging these sources, the text are cleaned by removing common digital artifacts such as hashtags, non-breaking space markers ("NBSP"), and citation patterns. The merged and cleaned text was then segmented into sentences. This process yielded over 400,000 clean Hausa sentences, distributed as follows: Hausa Wikipedia (160k), Wikimedia Hausa (214k), Other Texts (29k). The validation and training sets each contain 5000 sentences.

Because there is no readily available data annotated with errors for Hausa, we are facing with a chicken and the egg problem. As such the majority of Hausa errors are not to be found in (quasi-) offi-

cial sources such our dataset, but in more informal texts published on social media.

We use the NaijaSenti dataset of approximately 10,000 Hausa tweets (Muhammad et al., 2022, 2023), which typically contain the types of problems we are seeking to address. Potential misspellings are first identified using a lexicon-based approach: tokens absent from a standardized Hausa word list are flagged as out-of-vocabulary (OOV). For each OOV token, we compute its similarity to in-vocabulary items using a normalized Levenshtein distance. Tokens longer than two characters are grouped into length-based buckets $[\text{len}(x) - 1, \text{len}(x) + 1]$, and within each bucket, normalized Levenshtein distances are computed for all word pairs. Using DBSCAN (`eps`=0.4, `min_samples`=2) with the precomputed distance matrix, clusters of similar misspellings are identified while ignoring singleton (noise) clusters. From these results, we collect distributions of normalized edit distances and cluster sizes, which are visualized as histograms for both datasets (see Figure 1). We derive a distribution of error types, distinguishing between character substitutions insertions, deletions, and spacing errors. The types of noise introduced include: incorrect character substitution by replacing Hausa-specific hooked letters (ɓ, ɗ, ƙ, ƴ and their uppercase counterparts) with their plain English alphabet equivalents (b, d, k, y). Random spacing errors and space removal errors, e.g., merging adjacent words by removing the intervening space. Random character deletion, random character duplication, random substitution of some characters with other characters. Chunk deletion and insertion, operating on 2-character chunks from words. These changes are irreversible with a rule-based system. The probabilities of these operations are selected randomly at first, before being subsequently visualized against the normalized distances of the Twitter dataset. The process is repeated until the normalized distances of the synthetic dataset match that of the Twitter dataset. To assess the reliability of the synthetically generated noise, a native speaker manually inspects the generated content across different text genres in a shallow and non-exhaustive manner. The derived probabilities of these errors are rendered in Table 1 and are used to generate **synthetic alterations** to the original dataset to reflect human-like anomalies. Additionally, we compare the distribution of writing errors in the synthetic dataset using the Jensen–Shannon (JS) distance between two discrete distributions T (twitter) and O (originals), defined as

$$JS(T \,\|\, O) = \sqrt{\tfrac{1}{2} D_{KL}(T \,\|\, M) + \tfrac{1}{2} D_{KL}(O \,\|\, M)}$$

where $M = \frac{1}{2}(T + O)$ and D_{KL} denotes the Kullback–Leibler divergence. The computed JS distance of 0.14 indicating a good match between the empirical and synthetic distance distributions.

Noise Type	Probability
random_spacing	0.02
remove_spaces	0.15
incorrect_characters	0.02
delete_characters	0.005
duplicate_characters	0.01
substitute_characters	0.001
transpose_characters	0.01
delete_chunk	0.0015
insert_chunk	0.001

Table 1: Character-Level Noise Configuration.

4 Automatic Correction Models

The automatic correction task is framed as translating noisy Hausa sentences (input sequence) into its corrected standard form (target sequence). We investigate a variety of multilingual and African transformer models by fine-tuning them on the synthetically generated noisy-clean parallel Hausa dataset. We adopt the following pre-trained models that have been exposed to Hausa.

- **M2M100** (Fan et al., 2020) is a seq2seq massively multilingual model, we use the 418M parameter version. These models are designed for many-to-many multilingual translation and can be adapted for monolingual correction tasks.
- **AfriTeVA**: we utilize both the "base", "small" and "large" versions of AfriTeVA (Jude Ogundepo et al., 2022), a seq2seq model pretrained on a diverse set of African languages, including Hausa;
- **mT5** (Xue et al., 2021) is a multilingual T5 transformer including Hausa in its 101 pretraining languages.
- **AfriMBART** (Adelani et al., 2022) a multilingual BART model (Tang et al., 2020) trained for machine translation of 16 African languages

Model	BLEU ↑	METEOR ↑	F1 ↑	WER ↓	CER ↓
Copy Baseline	0.293	0.654	0.554	0.501	0.082
LSTM Baseline	0.156	0.389	0.757	0.635	0.534
GPT-5.1 Baseline	0.595	0.743	0.773	0.271	0.067
NCAIR1/N-ATLaS (8B)	0.886	0.929	0.936	0.079	0.013
M2M100 (418M)	0.853	0.932	0.933	0.079	0.017
UBC-NLP/cheetah-base (1.2B)	0.834	0.924	0.926	0.086	0.028
AfriTeVA-Base (580M)	0.754	0.880	0.884	0.135	0.045
AfriTeVA-Small (229M)	0.656	0.823	0.835	0.203	0.075
AfriMBART (680M)	0.789	0.853	0.855	1.862	1.825
mBART-Large-50 (610M)	0.413	0.661	0.730	0.352	0.090
mT5-Base (580M)	0.656	0.823	0.841	0.267	0.158

Table 2: Performance of various transformer models on the Hausa text correction test set. Best performing models are highlighted with gray. Lower is better for WER and CER; higher is better for others. The M2M100 adapted for the text normalization task achieves strong results, better than models multilingually adapted to African languages. The only model that is marginally better is a finetuned 8B model. GPT-5.1 under-performs for Hausa text normalization and is comparable to the baseline of doing no transformation to the data (first row).

- **UBC-NLP/cheetah-base** (Adebara et al., 2024) a massively multilingual T5 language model for 517 African languages and language varieties.
- **NCAIR1/N-ATLaS** (Technologies et al., 2025) is a fine-tuned multilingual language model based on Llama-3 8B designed to support African languages, including Hausa, Igbo, and Yoruba alongside English.

These models are fine-tuned using the noisy Hausa sentences as input and the corresponding clean sentences as the target output. The N-ATLaS model is fully finetuned for two epochs using Axolotl (Axolotl maintainers and contributors, 2023) framework and a base instruction copied verbatim in Appendix D. Additionally, we employ three more baselines: 1. a copy baseline that simply copies the noisy texts with no operation; 2. an LSTM-based sequence-to-sequence model; and 3. GPT-5.1 prompted three times (see Appendix C) with default parameters, average scores are reported ± 0.001. Details on hyperparameters for all the other models are available in Appendix A and in the released code.

4.1 Automatic Evaluation Metrics

Character Error Rate **(CER)** (Wigington et al., 2017; James et al., 2024) measures the minimum number of edit operations (insertions, deletions, substitutions) required to transform the predicted output into the reference text, divided by the length of the reference text. Word Error Rate **(WER)** calculates the number of word-level edits needed to transform the corrected sentence into the reference sentence, divided by the number of words in the reference sentence. The metric is useful when errors predominantly affect word boundaries or word choice (NithyaKalyani and Jothilakshmi, 2019). **BLEU**[2] from SacreBleu (Post, 2018) measures n-gram precision overlap between the generated text and the reference. **METEOR** - Metric for Evaluation of Translation with Explicit ORdering evaluates translations by aligning unigrams between the hypothesis and reference, computing a score based on precision, recall, and a fragmentation penalty. Token **F1-Score** calculates the score based on the overlap of tokens between the predicted and reference sentences. It provides a balanced measure of token-level precision and recall.

5 Results

Model performance is presented in Table 2, for BLEU, METEOR, and F1, higher scores are better, while for WER and CER, lower scores indicate better performance. The results demonstrate that Hausa text correction is an achievable task, with the M2M100 (418M) model obtaining the best scores across all metrics, significantly restoring the noisy text to the original quality.

Surprisingly, M2M100 performs better than mod-

[2]Signature: `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.4.3`

Model	BLEU ↑	METEOR ↑	F1 ↑	WER ↓	CER ↓
NCAIR1/N-ATLaS (8B)	0.808	0.879	0.892	0.169	0.089
M2M100 (418M)	0.792	0.885	0.896	0.117	0.024
UBC-NLP/cheetah-base	0.740	0.860	0.871	0.174	0.080
AfriTeVA-Base (580M)	0.640	0.817	0.847	0.625	0.512
AfriTeVA-Small (229M)	0.547	0.770	0.820	1.039	0.874
AfriMBART (680M)	0.632	0.714	0.720	5.991	6.423

Table 3: Evaluation results on Hausa noisy-to-clean correction for out-of-domain data. The values are overall smaller than for the in-domain data Table 2, however the hierarchy of the models is preserved - M2M being the best performing model followed by Cheetah.

els such as AfriTeVA, `UBC-NLP/cheetah-base` and other models exposed to African languages. The model is on par with a much larger and more powerful model based on LLaMa 3 8B that has been instruction tuned on over 400 million tokens of multilingual instruction data for African languages – `NCAIR1/N-ATLaS`. The latter achieves the best overall scores (± 0.001 during multiple runs), however, given the much larger compute requirements of this model, we consider the best-performing model to be the smaller and more efficient M2M100.

Our experiments indicate that both the pretraining strategy and the architecture are important. AfriTeVA is a generic T5-like architecture pretrained on 10 African languages of which three languages dominate the corpus Swahili, Tigrinya, and Hausa, languages that are from different language families: Bantu, Cushitic, Chadic. Cheetah is also a T5-like model, while M2M100 is an encoder–decoder model pretrained on a task much more similar to ours: many-to-many translations in 100 languages, including 18 African languages.

Furthermore, there is also the possibility of data contamination. The M2M100 is pre-trained on data that is not public, which, according to Fan et al. (2020), the majority of data is Common Crawl and synthetic translations. The training of AfriTeVA is the AfriBERTa corpus (Ogueji et al., 2021) which is also based on Common Crawl with more focus on quality. To test the data contamination hypothesis, all our evaluations exclude MinHashLSH approximate overlaps (Zhu et al., 2024) between AfriBERTa and our test set. We could only find 56 such overlaps at a 0.5 Jaccard coefficient threshold and their impact amounts in less than 0.001 changes in the final metrics. Additionally, we evaluate the models exclusively on the private OOD dataset (see Table 3), which reveals an identical performance hierarchy among the models, albeit with lower overall scores.

Unsurprisingly, models lacking explicit Hausa pre-training such as mT5 and mBART perform subpar to any of the models pre-trained on Hausa and African languages, with scores similar to GPT-5.1 and close to the Copy baseline.

5.1 Human Evaluation

We conduct a human evaluation of the best model's predictions using a set of 500 randomly selected genre-diverse samples. A native Hausa speaker is asked to provide binary assessments of correctness for the automatically corrected texts. The evaluation reveals that 358 predictions (71.6%) are correct according to the reference. Within these correct predictions, 57 samples (12.5%) contain at least one instance where hooked Hausa characters (ɓ, ɗ, ƙ, ƴ) are swapped with their Latin-looking counterparts. Furthermore, 135 samples (27%) are labeled as incorrect. This incorrect label comprises instances where the model's output did not precisely match the reference word for word but the core meaning remained largely intact. Only a small fraction of 7 sentences (1.4%) out of the total 500, were judged to have predictions that were worse than the original input.

A manual comparison between the M2M models and N-ATLaS model reveals a very strong tie, henceforth, the human evaluation was focused more on AfriTeVA. These models, due to their pretraining task, tend to reconstruct the errors with paraphrasing and different words from the reference. For example, the model predicts "... wasannin gudun mita 100" (*100m running competitions*) instead of the true "... tseren gudun mita 100" (*100m race*). Semantically, the sentences may be similar, but this behavior is a source of downstream errors.

Appendix B contains several examples of predicted texts from the best-performing model. The

model tends to preserve the overall coherence of the phrase. As such, typos or misspellings that do not affect the coherence (i.e., a change of tense / person), but alter the original meaning are not modified. Furthermore, since the base model is multilingual, we observe several cases where named entities are contaminated from other languages - these are not corrected or transcribed using the standards for Hausa (e.g.,*Pridnestrovie*), but using alternative spellings that are valid in other languages (e.g., *Pridnistria*).

6 Impact on Downstream Tasks

To assess the practical impact of text correction, we conduct downstream evaluations on several tasks: text genre detection, machine translation, and LLM prompt response. These evaluations are designed to identify the extent to which noisy data degrades performance across tasks relative to the original data, and to determine whether automatic correction can recover the original performance levels.

6.1 Genre Detection

For this task we use the existing 750 chunks of the MultiplEYE out-of-domain dataset, translated and edited into Hausa by a professional translator.

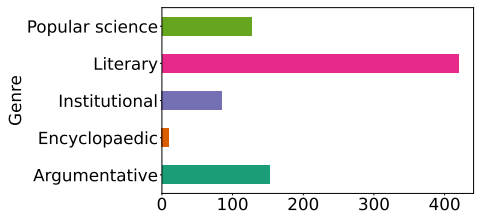

Figure 2: Distribution of genres in the data.

The data contains two argumentative texts from Programme for International Student Assessment (PISA), one text from Wikipedia, a part of the Universal Declaration of Human Rights, a text from the European Commission, seven literary texts, and two pop-science texts. The majority class consists in the Literary genre, as seen in Figure 2 and we do stratified random split for model evaluation.

We train a genre-detection classifier at the sentence level based on the AfriBERTa-base[3] (Ogueji et al., 2021) model and evaluate the model using stratified 5-fold cross validation. The comparative evaluation scores between classifiers trained on noise-induced data and automatically cleaned data and report the results in Table 4. **Original** represents human-edited, clean reference text (either from existing clean corpora or from professional translation). **Noisy** represents texts obtained by injecting synthetic noise into the original text. **Cleaned** are automatically corrected text produced by the normalization model when applied to the noisy input.

Data	Accuracy	F1
Original	80% ± 2%	0.80 ± 0.02
Noisy	78% ± 1%	0.77 ± 0.01
Cleaned	81% ± 3%	0.80 ± 0.03

Table 4: Genre detection performance comparison between models trained on noisy data versus cleaned data. Accuracy is slightly decreased on noised data while automatic correction recovers the original performance.

The noise in the data reduces the ability of the classifiers to correctly predict the genre while the automatic cleaning restores the ability of models to learn and make predictions from the data, albeit with a slight increase in standard deviation.

We have experimented on the `mteb/NaijaSenti` sentiment dataset as well, without observing any statistically significant difference between the model on the noisy data vs. the model on the corrected dataset (78% accuracy in both cases). The impact of noise on text classification is less stringent compared to other tasks, as we will see further.

6.2 Hausa-to-English Machine Translation Task

Translation quality is evaluated using a subset of the FLORES+ benchmark (NLLB Team et al., 2024; Abdulmumin et al., 2024) consisting of 997 sentence human translated English-Hausa pairs. Noise was added using the same procedure as for the rest of the data using the parameters in Table 1.

To translate, we use two proprietary models: Google Translate and GPT-5.1, and two open source models: NLLB-200 (NLLB Team et al., 2024) and Tencent Hunyuan 7B (Zheng et al., 2025). The latter obtained the best evaluation scores across 30 out of 31 languages at the recent WMT2025 (Kocmi et al., 2025). For LLM-based translation, the prompts used for the LLM-based translation systems are provided in Appendix E. The evaluation is carried using string based-metric BLEU with and MetricX-24-Hybrid-XL (Juraska et al., 2024) and the results are rendered in Table 5.

[3] https://huggingface.co/castorini/afriberta_base

System	Metric	Original	Noisy	Cleaned
GPT-5.1	BLEU	33.3	24.47	27.26
	MetricX	0.812	0.689	0.717
Google Trans.	BLEU	42.01	33.07	35.02
	MetricX	0.816	0.701	0.722
NLLB-200	BLEU	36.46	26.88	30.54
	MetricX	0.762	0.596	0.642
Hunyuan	BLEU	13.58	7.9	10.34
	MetricX	0.5	0.382	0.447

Table 5: Translation quality comparison on original, noisy, and cleaned Hausa texts across four translation systems on the FLORES+ dataset. The automatically corrected texts consistently obtain better translations than the noisy texts.

The best translation model, or rather the one that matches the best the set FLORES+ is Google Translate, followed by the open-source NLLB-200, GPT-5.1 zero shot prompting, and Hunyuan MT. The latter has not been exposed to Hausa, therefore it is expected to perform worse.

A manual analysis of translations based on their similarity indicates that translations from cleaned texts are generally very similar to those from original texts, rather than closer to the reference. It is expected to observe lower evaluation scores of the cleaned sentences because the cleaned text is produced automatically and, while it removes orthographic noise, it may introduce unintended changes (see examples in Appendix B). There are also artifacts in the FLORES+ dataset where the references are not a literal translation or cases where there is not enough context to translate properly the original (e.g., rows 226, 334, 495 in the development split[4]).

For machine translation, unlike the text classification task, automatic recovery of errors is not perfect and phrases can lose their initial meaning, therefore being mistranslated into English.

6.3 Instruction Following

This task is tested using a set of benchmarking data for Hausa that evaluate several tasks such as Named Entity Recognition (MasakhaNER) (Adelani et al., 2021) and the IrokoBench datasets recently published by Adelani et al. (2025b) comprising of Mathematical Reasoning QA (AfriMGSM), Natural Language Inference (AfriXNLI), and Multi-Choice QA (AfriMMLU). The latter is split by subject, such as elementary mathematics, global facts, high school geography, high school microeconomics, and international law.

[4] https://huggingface.co/datasets/openlanguagedata/flores_plus

Task	Original	Noisy	Cleaned
NER	0.64	0.19	0.37
AfriMMLU (all)	79	74	76
Elem. Math	82	79	78
Facts	72	68	68
Geography	82	73	77
Economics	78	75	78
Law	79	78	82
AfriMGSM	58	49	49
AfriXNLI	73	64	66

Table 6: Performance comparison across Original, Noisy, and Cleaned inputs. Named Entity Recognition is evaluated using the F1 score while all the others using accuracy. NER is the most impacted task because it depends on identifying the exact word boundary of entities.

We use the GPT-5.4 proprietary model and evaluate the zero-shot performance over the original, noised, and cleaned data, acknowledging that proprietary models may likely be contaminated by the data, especially since the benchmarks are translations of English variants. We evaluate NER using F1 score and all the other tasks using plain accuracy.

Table 6 shows that proprietary large language models are sensitive to the noise introduced in the prompts and that a pre-processing step of cleaning and normalizing the data can lead to better instruction-following abilities. The named entity recognition (NER) task is the most affected (dropping from 0.64 original F1 score to 0.19 with noise and 0.37 cleaned) because the word boundaries are not preserved perfectly after cleaning up the noise. Despite this drop, the improvements over the noisy data is almost double.

For multiple answer questions, i.e., Measuring Massive Multitask Language Understanding (MMLU), each subject has 100 questions and we can observe variation across subjects; for instance, geography and facts are more impacted than law or economics. Some of the results are incidental, such the law, where the cleaned version obtains higher scores than all the others, a result that brings up some limitations of this type of evaluative comparison. Similarly, the drop in performance for Natural Language Understanding (NLI) ($73 \rightarrow 64 \rightarrow 66$) shows that text normalization is not able to restore the quality in order for commercial models to predict the correct textual entailment and that there is

a surprisingly close gap between noised and corrected texts.

The African Multilingual Grade School Math Benchmark (MGSM) exhibits a substantial performance drop under noisy conditions, which is not recovered after cleaning.

Overall, the findings indicate that noise robustness varies considerably by task, with structured prediction tasks being most vulnerable, and that simple cleaning methods provide uneven gains depending on the nature of the task.

7 Conclusions

This paper addresses the prevalent issue of orthographic writing anomalies in digital Hausa text, which impedes the development of robust NLP tools.

We compare several solutions centered on fine-tuning transformer models using a novel, synthetically-generated parallel corpus of over 400,000 clean-noisy Hausa sentence pairs. The dataset is designed to reflect realistic error patterns.

Our experiments demonstrate that a sequence-to-sequence M2M100 (418M) model achieves a performance comparable to 8B models supervised-fine-tuned on Hausa and other African languages. A human and automatic evaluation of the methods also reveals the limitations of normalization and that the models may not be able to recover entirely the original data, which prompts future directions to be investigated for Hausa.

Overall, these findings show that smaller fine-tuned models, trained on synthetically noised data, can substantially enhance Hausa text quality and a downstream NLP tasks including machine translation, text classification, and LLM instruction following. This has significant implications for Hausa NLP, as cleaner text resources are crucial for developing more accurate and reliable downstream applications like machine translation and sentiment analysis, thereby improving digital resources for the Hausa-speaking community. This work provides a step towards enhancing Hausa language resources and offers a replicable methodology for tackling similar challenges in other low-resource languages.

To ensure full reproducibility and facilitate further research, the synthetically generated noisy-clean parallel dataset, along with the code for our models, has been made publicly available.[5]

[5]https://github.com/ahmadmwali/HausaSeq2Seq

Limitations

The current limitations of our work that we have identified consist of

1. the usage of social media data as a source of noise and user-generated anomalies; this noise is then projected onto good-quality edited texts

2. the edited texts are from different genres and the synthetic noise might not be naturally occurring in such documents

3. the evaluation of downstream tasks uses only automatic metrics which might introduce particular biases towards certain languages, further analysis is required to understand which parts of the text are irreparable and whether this is an actual issue in online texts

Acknowledgments

We would like to express our gratitude to Habib Sani Yahaya for his contribution in translating and revising the MultiplEYE texts.

This research is supported by InstRead: Research Instruments for the Text Complexity, Simplification and Readability Assessment CNCS - UEFISCDI project number PN-IV-P2-2.1-TE-2023-2007 and by the project "Romanian Hub for Artificial Intelligence - HRIA", Smart Growth, Digitization and Financial Instruments Program, 2021-2027, MySMIS no. 351416.

References

Idris Abdulmumin, Sthembiso Mkhwanazi, Mahlatse S. Mbooi, Shamsuddeen Hassan Muhammad, Ibrahim Said Ahmad, Neo N. Putini, Miehleketo Mathebula, Matimba Shingange, Tajuddeen Gwadabe, and Vukosi Marivate. 2024. Correcting FLORES evaluation dataset for four African languages. In *Proceedings of the Ninth Conference on Machine Translation*, Miami, USA. Association for Computational Linguistics.

Amina Imam Abubakar, Abubakar Roko, AB Muhammad, and Ibrahim Saidu. 2019. Hausa wordnet: An electronic lexical resource. *Saudi Journal of Engineering and Technology*, 4(8):279–285.

Ife Adebara, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2024. Cheetah: Natural language generation for 517 African languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12798–12823, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, and 26 others. 2022. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, and 42 others. 2021. MasakhaNER: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

David Ifeoluwa Adelani, A. Seza Doğruöz, Iyanuoluwa Shode, and Anuoluwapo Aremu. 2025a. Does generative AI speak Nigerian-Pidgin?: Issues about representativeness and bias for multilingualism in LLMs. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1571–1583, Albuquerque, New Mexico. Association for Computational Linguistics.

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba Oluwadara Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Ijeoma Chukwuneke, Happy Buzaaba, Blessing Kudzaishe Sibanda, Godson Koffi Kalipe, Jonathan Mukiibi, Salomon Kabongo Kabenamualu, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndolela, and 8 others. 2025b. IrokoBench: A new benchmark for African languages in the age of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2732–2757, Albuquerque, New Mexico. Association for Computational Linguistics.

Ibrahim Ahmad, Shiran Dudy, Resmi Ramachandranpillai, and Kenneth Church. 2024. Are generative language models multicultural? a study on Hausa culture and emotions using ChatGPT. In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 98–106, Bangkok, Thailand. Association for Computational Linguistics.

Ibrahim Said Ahmad, Shiran Dudy, Tadesse Destaw Belay, Idris Abdulmumin, Seid Muhie Yimam, Shamsuddeen Hassan Muhammad, and Kenneth Church. 2025. Exploring cultural nuances in emotion perception across 15 african languages. *Preprint*, arXiv:2503.19642.

Adewale Akinfaderin. 2020. HausaMT v1.0: Towards English–Hausa neural machine translation. In *Proceedings of the Fourth Widening Natural Language Processing Workshop*, pages 144–147, Seattle, USA. Association for Computational Linguistics.

Jesujoba Oluwadara Alabi, Israel Abebe Azime, Miaoran Zhang, Cristina España-Bonet, Rachel Bawden, Dawei Zhu, David Ifeoluwa Adelani, Clement Oyeleke Odoje, Idris Akinade, Iffat Maab, Davis David, Shamsuddeen Hassan Muhammad, Neo Putini, David O. Ademuyiwa, Andrew Caines, and Dietrich Klakow. 2025. AFRIDOC-MT: Document-level MT corpus for African languages. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27770–27806, Suzhou, China. Association for Computational Linguistics.

Axolotl maintainers and contributors. 2023. Axolotl: Open source llm post-training.

Muazzam Bashir, Azilawati Rozaimee, and Wan Malini Wan Isa. 2017. Automatic hausa language-text summarization based on feature extraction using naïve bayes model. *World Applied Science Journal*, 35(9):2074–2080.

Tadesse Destaw Belay, Israel Abebe Azime, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Idris Abdulmumin, Abinew Ali Ayele, Shamsuddeen Hassan Muhammad, and Seid Muhie Yimam. 2025. AfroXLMR-social: Adapting pre-trained language models for African languages social media text. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15570–15587, Suzhou, China. Association for Computational Linguistics.

Roger Blench. 2012. *An atlas of Nigerian languages*. Kay Williamson Educational Foundation Oxford.

Everlyn Asiko Chimoto, Constantine Lignos, Shamsuddeen Muhammad, Idris Abdulmumin, Clemencia Siro, and David Ifeoluwa Adelani, editors. 2026. *Proceedings of the 7th Workshop on African Natural Language Processing (AfricaNLP 2026)*. Association for Computational Linguistics, Rabat, Morocco.

Berthold Crysmann. 2016. Representing morphological tone in a computational grammar of hausa. *Journal of Language Modelling*, 3(2).

Cheikh M Bamba Dione, David Ifeoluwa Adelani, Peter Nabende, Jesujoba Alabi, Thapelo Sindane, Happy Buzaaba, Shamsuddeen Hassan Muhammad, Chris Chinenye Emezue, Perez Ogayo, Anuoluwapo Aremu, and 1 others. 2023. Masakhapos: Part-of-speech tagging for typologically diverse african languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10883–10900.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2023. Ethnologue: Languages of the world.

Chris Emezue, Hellina Nigatu, Cynthia Thinwa, Helper Zhou, Shamsuddeen Muhammad, Lerato Louis, Idris Abdulmumin, Samuel Oyerinde, Benjamin Ajibade, Olanrewaju Samuel, Oviawe Joshua, Emeka Onwuegbuzia, Handel Emezue, Ifeoluwatayo A. Ige, Atnafu Lambebo Tonja, Chiamaka Chukwuneke, Bonaventure F. P. Dossou, Naome A. Etori, Mbonu Chinedu Emmanuel, and 3 others. 2023. The african stopwords project: curating stopwords for african languages. *Preprint*, arXiv:2304.12155.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *Preprint*, arXiv:2010.11125.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Nora Hollenstein, Marie-Luise Müller, Deborah N. Jakobi, Cui Ding, Maja Stegenwallner-Schütz, Ana Matić, Eva Pavlinušić Vilus, Ramunė Kasperė, Anna Bondar, Maroš Filip, Stefan Frank, Jana Hofmann, Thyra Krosness, Kaidi Lõo, Johanne Nedergaard, Chiara Tschirner, and Lena A. Jäger. 2026. *MultiplEYE Data Collection Guidelines*.

Kedir Yassin Hussen, Walelign Tewabe Sewunetie, Abinew Ali Ayele, Sukairaj Hafiz Imam, Shamsuddeen Hassan Muhammad, and Seid Muhie Yimam. 2025. The state of large language models for african languages: Progress and challenges. *Preprint*, arXiv:2506.02280.

Aizhan Imankulova, Takayuki Sato, and Mamoru Komachi. 2019. Filtered pseudo-parallel corpus improves low-resource neural machine translation. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 19(2):1–16.

Jesin James, Deepa P Gopinath, and 1 others. 2024. Advocating character error rate for multilingual asr evaluation. *arXiv preprint arXiv:2410.07400*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world.

Odunayo Jude Ogundepo, Akintunde Oladipo, Mofetoluwa Adeyemi, Kelechi Ogueji, and Jimmy Lin. 2022. AfriTeVA: Extending ?small data? pretraining approaches to sequence-to-sequence models. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 126–135, Hybrid. Association for Computational Linguistics.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Alina Karakanta, Jon Dehdari, and Josef Van Genabith. 2017. Neural machine translation for low-resource languages without parallel corpora. *Machine Translation*, 32(1–2):167–189.

Ramunė Kasperė, Anna Bondar, Sergiu Nisioi, Maja Stegenwallner-Schütz, Hanne B. Søndergaard Knudsen, Ana Matić, Eva Pavlinušić Vilus, Dorota Klimek-Jankowska, Chiara Tschirner, Not Battesta Soliva, Deborah N. Jakobi, Cui Ding, Dima Abu Romi, Cengiz Acarturk, Matilda Agdler, Anton Marius Alexandru, Mohd Faizan Ansari, Annalisa Arcidiacono, Elizabete Ausma Velta Barisa, and 87 others. 2026. The multipleye text corpus: Towards a diverse and ever-expanding multilingual text corpus. In *Proceedings of the 2026 International Conference on Language Resources and Evaluation (LREC 2026)*, Rabat, Morocco. European Language Resources Association and International Committee on Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Natalia Fedorova, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Kenton Murray, Masaaki Nagata, and 9 others. 2025. Preliminary ranking of wmt25 general machine translation systems. *Preprint*, arXiv:2508.14909.

Constantine Lignos, Idris Abdulmumin, and David Adelani, editors. 2025. *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*. Association for Computational Linguistics, Vienna, Austria.

Stefano Lusito, Edoardo Ferrante, and Jean Maillard. 2023. Text normalization for low-resource languages: the case of ligurian. In *Proceedings of the sixth workshop on the use of computational methods in the study of endangered languages*, pages 98–103.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Felermino D'ario M'ario Ant'onio Ali, Davis Davis, Sa-

lomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Gwadabe, Samuel Rutunda, and 7 others. 2023. Afrisenti: A twitter sentiment analysis benchmark for african languages.

Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Sebastian Ruder, Ibrahim Sa'id Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Salahudeen Abdullahi, Anuoluwapo Aremu, Alípio Jorge, and Pavel Brazdil. 2022. NaijaSenti: A Nigerian Twitter sentiment corpus for multilingual sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 590–602, Marseille, France. European Language Resources Association.

Shamsuddeen Hassan Muhammad, Ibrahim Said Ahmad, Idris Abdulmumin, Falalu Ibrahim Lawan, Sukairaj Hafiz Imam, Yusuf Aliyu, Sani Abdullahi Sani, Ali Usman Umar, Tajuddeen Gwadabe, Kenneth Church, and Vukosi Marivate. 2025. HausaNLP: Current status, challenges and future directions for Hausa natural language processing. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pages 176–191, Vienna, Austria. Association for Computational Linguistics.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Tajudeen Kolawole, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Hassan Muhammad, Salomon Kabongo, Salomey Osei, Sackey Freshia, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa, Mofe Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, and 29 others. 2020. Participatory research for low-resourced machine translation: A case study in african languages. *arXiv (Cornell University)*.

Paul Newman. 2000. *The Hausa Language: An Encyclopedic Reference Grammar*. Yale University Press, New Haven.

S. Nirenburg. 2009. *Language Engineering for Lesser-Studied Languages*. IOS Press.

Sergiu Nisioi, Anna Bondar, Ramunė Kasperė, and Maja Stegenwallner-Schütz. 2026. The multipleye text corpus data and materials.

A. NithyaKalyani and S. Jothilakshmi. 2019. *Speech Summarization for Tamil Language*, pages 113–138. Academic Press.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

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.

Arturo Oncevay, Gerardo Cardoso, Carlo Alva, César Lara Ávila, Jovita Vásquez Balarezo, Saúl Escobar Rodríguez, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Juan López Bautista, Nimia Acho Rios, Remigio Zapata Cesareo, Héctor Erasmo Gómez Montoya, and Roberto Zariquiey. 2022. SchAman: Spell-checking resources and benchmark for endangered languages from amazonia. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 411–417, Online only. Association for Computational Linguistics.

Shantipriya Parida, Idris Abdulmumin, Shamsuddeen Hassan Muhammad, Aneesh Bose, Guneet Singh Kohli, Ibrahim Said Ahmad, Ketan Kotwal, Sayan Deb Sarkar, Ondřej Bojar, and Habeebah Kakudi. 2023. HaVQA: A dataset for visual question answering and multimodal research in Hausa language. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10162–10183, Toronto, Canada. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Ochilbek Rakhmanov and Tim Schlippe. 2022. Sentiment analysis for Hausa: Classifying students' comments. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 98–105, Marseille, France. European Language Resources Association.

Saheed Abdullahi Salahudeen, Falalu Ibrahim Lawan, Ahmad Wali, Amina Abubakar Imam, Aliyu Rabiu Shuaibu, Aliyu Yusuf, Nur Bala Rabiu, Musa Bello, Shamsuddeen Umaru Adamu, and Saminu Mohammad Aliyu. 2023. Hausanlp at semeval-2023 task 12: Leveraging african low resource tweetdata for sentiment analysis. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 50–57.

Sani Abdullahi Sani, Shamsuddeen Hassan Muhammad, and Devon Jarvis. 2025. Investigating the impact of language-adaptive fine-tuning on sentiment analysis in Hausa language using AfriBERTa. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 101–111, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Iyanuoluwa Shode, David Ifeoluwa Adelani, Jing Peng, and Anna Feldman. 2023. Nollysenti: Leveraging transfer learning and machine translation for nigerian movie sentiment classification. *arXiv (Cornell University)*.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

Awarri Technologies, National Information Technology, and Development Agency. 2025. N-atlas-llm: A multilingual african language model. Fine-tuned Llama-3 8B model for African languages developed in collaboration with the Federal Government of Nigeria.

Kosei Uemura, Miaoran Zhang, and David Ifeoluwa Adelani. 2026. AfriMTEB and AfriE5: Benchmarking and adapting text embedding models for African languages. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3697–3717, Rabat, Morocco. Association for Computational Linguistics.

Aghaegbuna Uwaezuoke and Gift Anachunam. 2023. A contrastive study of reduplication in the igbo and hausa languages. *Journal of The Linguistic Association of Nigeria*, 26(1):82–105.

Adam Wiemerslage, Miikka Silfverberg, Changbing Yang, Arya McCarthy, Garrett Nicolai, Eliana Colunga, and Katharina Kann. 2022. Morphological processing of low-resource languages: Where we are and what's next. *Findings of the Association for Computational Linguistics: ACL 2022*, pages 988–1007.

Curtis Wigington, Seth Stewart, Brian Davis, Bill Barrett, Brian Price, and Scott Cohen. 2017. Data augmentation for recognition of handwritten words and lines using a cnn-lstm network. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, pages 639–645.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Rufai Yusuf Zakari, Zaharaddeen Karami Lawal, and Idris Abdulmumin. 2021. A systematic literature review of hausa natural language processing. *International Journal of Computer and Information Technology*, 10(4).

Mao Zheng, Zheng Li, Bingxin Qu, Mingyang Song, Yang Du, Mingrui Sun, and Di Wang. 2025. Hunyuan-mt technical report. *Preprint*, arXiv:2509.05209.

Eric Zhu, Vadim Markovtsev, Aleksey Astafiev, Arham Khan, Chris Ha, Wojciech Łukasiewicz, Adam Foster, Sinusoidal36, Spandan Thakur, Stefano Ortolani, Titusz, Vojtech Letal, Zac Bentley, fpug, hguhlich, long2ice, oisincar, Ron Assa, Senad Ibraimoski, and 8 others. 2024. ekzhu/datasketch: v1.6.5.

A Training Parameters

Hyperparameters for the T5, BART, and M2M models. Default training hyperparameters include a learning rate of 2e-5, a training batch size of 4 per device, an evaluation batch size of 8 per device, 3 training epochs, gradient accumulation steps of 4, and weight decay of 1e-3. The maximum sequence length for tokenization was set to 256 tokens, and generation during evaluation was configured with a maximum of 256 new tokens and 4 beams for beam search. Specific LoRA configurations involved a rank (r) of 16, alpha of 32, and dropout of 0.05. Training was performed on NVIDIA L4 and H100 GPUs with the *accelerate* library.

Hyperparameters for LLM fine-tuning The experiments use the Axolotl framework initialized from the `NCAIR1/N-ATLaS`. Training uses an instruction-formatted Hausa corpus derived from the *noisy*-to-*clean* correction task, reserving 5% of the training file for validation. Model updates are performed with a micro-batch size of 1 and gradient accumulation steps of 32. The optimizer is an 8-bit AdamW from BitsAndBytes, configured with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$, and a weight decay of 0.001. The learning rate was set to 2×10^{-5} using a cosine learning-rate scheduler with 100 warmup steps. Models were trained for two epochs, with evaluation conducted using batches of 8 samples and typical causal LM metrics (perplexity).

B Data Samples

Below we provide several examples of sentences that are automatically standardized using our best model. We have selected both perfect reconstructions and sentences with different types of errors. With boldface we highlight the main errors that make the prediction different from the true reference.

1. **Input (Noisy):** AnhaifiAbubakar Malami ,musulmin fulani , a ranar 17 gaAfrilunshekara ta 1967 a Birnin Kebbi, babban birnin JiharKebbi, Arewacin Najeria y .
   Prediction: An haifi Abubakar Malami , musulmin fulani , a ranar 17 ga Afrilun shekara ta 1967 a Birnin Kebbi , babban birnin Jihar Kebbi , Arewacin Najeriya .
   Reference: An haifi Abubakar Malami , musulmin fulani , a ranar 17 ga Afrilun shekara ta 1967 a Birnin Kebbi , babban birnin Jihar Kebbi , Arewacin Najeriya .
   English translaiton: Abubakar Malami, a Fulani Muslim, was born on 17th April, 1967, in Birnin Kebbi, the Kebbi State capital, Northern Nigeria.
   Obs: The prediction is identical to the reference, the original meaning is preserved.

2. **Input (Noisy):** Transnistria (Prid nsetrovie) , kuma anafurta Transdniestria , shi ne wani ɓangare na Moldova gabashin koginDniesterr kuma (tun 1990)a ka-ayyanadakuma fiyeko ƙasa da aiki zaman kanta a jihar ba ta re dda wani kasada kasa fitarwa da ga duk wanisarki jihar .
   Prediction: Transnistria (**Pridnistria**) , kuma ana furta **Transnistria** , shi ne wani ɓangare na Moldova gabashin kogin Dniester kuma (tun 1990) **aka-yayya** da kuma fiye ko ƙasa da aiki zaman kanta a jihar ba tare da wani kasa da kasa fitarwa daga duk wani sarki jihar .
   Reference: Transnistria (Pridnestrovie) , kuma ana furta Transdniestria , shi ne wani ɓangare na Moldova gabashin kogin Dniester kuma (tun 1990) a ka-ayyana da kuma fiye ko ƙasa da aiki zaman kanta a jihar ba tare da wani kasa da kasa fitarwa daga duk wani sarki jihar .
   English translation: Transnistria (Pridnestrovie), also pronounced Transdniestria, is a part of Moldova east of the Dniester River and (since 1990) has been declared as, more or less, an independent state without any international recognition from any sovereign state.
   Obs: The models makes mistakes at the named entity level, it uses a (valid) multilingual replacement (*Pridnistria* from Macedonian) instead of the Hausa-specific variant (*Pridnestrovie*).

3. **Input (Noisy):** Kwararrunna'ura nadan adam (HMI) shine tsarin sarrafa allo nadijital (ta mafahii da fasahar Ha ptic) wanda kuma a karaba tare ad SF90Stradale.
   Prediction: Kwararrun na 'ura na dan adam (HMI) shine tsarin sarrafa allo na dijital (**ta mafi fiye da fasahar Hapttic**) wanda kuma aka **karaba** tare da SF90 Stradale .
   Reference: Kwararrun na'ura na dan adam (HMI) shine tsarin sarrafa allo na dijital (ta amfani da fasahar Haptic) wanda kuma aka raba tare da SF90 Stradale.
   English translation: Human Machine Interface (HMI) is a digital display control system (using Haptic technology) that is also shared with the SF90 Stradale.
   Obs: The original meaning is preserved, the model introduces a change of tense and a superlative expression.

4. **Input (Noisy):** Dukena Courland ya kaf a sansaninSt.Andrea indaya sayi bayi kum aya say da kayan Ingilishi ,amma a ciki n Fabrairu 1660 ya sayar da wurin ga Dutch .
   Prediction: Duk ne na Courland ya kafa sansanin St. Andrea inda ya sayi bayi kuma ya sayar da kayan Ingilishi , amma a cikin Fabrairu 1660 ya sayar da wurin ga Dutch .
   Reference: Duke na Courland ya kafa sansanin St. Andrea inda ya sayi bayi kuma ya sayar da kayan Ingilishi , amma a cikin Fabrairu 1660 ya sayar da wurin ga Dutch .
   English translation: The Duke of Courland established the St. Andrea fort where he bought slaves and sold English goods, but in February 1660 he sold the place to the Dutch.
   Obs: The model changes the subject of the sentence ("Duke of Courland" with "All of Courland", mistaking "Duke" with the Hausa word "duk" that translates to "all") making the sentence ungrammatical and difficult to read.

5. **Input (Noisy):** Hydrogen chloride, acetic acid , da mai yawansauran Brønsted-Lowry aci d ba zasu iya samar da hadin kai tare da nau 'in electron ba , duk da haka , don haka ba Lewis acid ba ne .
   Prediction: Hy (Sin , acetic acid , da mai yawan wasa na (n kungiya-L purposery aci da ba za su iya samar da hadin kai tare da nau 'in , duk da haka , don ba haka ba acid ba ne .
   Reference: Hydrogen chloride , acetic acid , da mafi yawan sauran Brønsted-Lowry acid ba za su iya samar da hadin kai tare da nau 'in electron ba , duk da haka , don haka ba Lewis acid ba ne .
   English translation: Hydrogen chloride, acetic acid, and most other Brønsted-Lowry acids cannot form coordination bonds with electron pairs, therefore, they are not Lewis acids.
   Obs: The meaning has changed from the original. Although the input is quite understandable, the model's prediction becomes completely incoherent (the model's prediction roughly translates to "Hy (China, acetic acid, and the very playful of the (n L-team purposery to eat with the possibility of providing cooperation together with the type, even though, therefore it is not an acid.")

C GPT Zero Shot Correction Prompt

You will be given a sentence in Hausa and your task is to automatically correct it. Please respect the following indications:

1. Read the sentence carefully, analyze it to understand it's context.
2. Correct any grammatical, spelling, or punctuation errors in the sentence.
3. Ensure that the corrected sentence maintains the original meaning and context.
4. Provide the corrected sentence as your output without any additional explanations or comments.
5. Do not change the structure of the sentence more than necessary to correct the errors.

D LLM SFT Instruction

Instruction:

Correct the noisy Hausa text below into a clean, grammatically correct version. Do not change the structure of the sentence.

Input:

{noisy}

Response:

E LLM Translation Prompt

You are a professional translator specializing in Hausa to English translation. You will receive text in Hausa that may contain informal language, colloquialisms, and errors.

Your task:

1. Translate the provided Hausa text to English accurately
2. Preserve the meaning and tone of the original text
3. For "noisy input": Translate as-is, maintaining any informal language or errors in meaning but translate them to English
4. For "clean input": Translate the clean, corrected version naturally

Guidelines:

- Maintain natural English flow while staying faithful to the Hausa original
- Keep cultural context and idioms where possible
- If text contains code-switching, translate only the Hausa portions
- Do not add explanations or notes, only provide the translation
- Preserve any numbers, names, or specific terms as they appear