FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 10 No. 12, June, 2026, pp 53 – 58
DOI: https://doi.org/10.33003/fjs-2026-1012-5226

An LLM-Driven Framework for Addressing Code-Switching and Orthographic Variance in Nigerian Language Data Collection

Rosemary Agharase Usiobaifo, *Roseline Oghogho Osaseri and Kaizer John Onyekachukwu

Department of Computer, Faculty of Physical Sciences, University of Benin, Edo State, Nigeria.

*Corresponding authors' email: roseline.osaseri@uniben.edu

ABSTRACT
Natural Language Processing (NLP) systems for low-resource languages continue to face significant challenges due to the poor quality and structural inconsistency of available textual data. In the Nigerian linguistic context, informal digital communication is characterized by frequent code-switching between English, Nigerian Pidgin, and indigenous languages, as well as high levels of orthographic variation. These characteristics reduce the effectiveness of conventional preprocessing pipelines, which are primarily designed for monolingual and standardized text. Existing approaches based on rule-based filtering, statistical methods, or conventional language identification models often fail to accurately interpret multilingual and inconsistent language patterns. This paper presents a reasoning-driven preprocessing framework for improving Nigerian language data collection for NLP systems. The proposed approach leverages Large Language Models (LLMs) to perform context-aware language identification, relevance filtering, and orthographic normalization. Unlike traditional preprocessing methods that rely mainly on lexical or statistical matching, the framework treats language identification and normalization as contextual reasoning tasks, preserving semantic meaning while reducing spelling variability and dataset noise. To evaluate the framework, textual data was collected from informal online sources, including blogs, forums, and social media platforms containing multilingual Nigerian language content. Experimental results demonstrate that the proposed approach achieved an 85% reduction in noisy data while improving the handling of code-switched and orthographically inconsistent text. The findings demonstrate the potential of LLM-based preprocessing to enhance data quality in low-resource NLP environments and provide a scalable foundation for developing more robust and inclusive language technologies for Nigerian languages.

Received: 05 June 2026
Accepted: 02 July 2026
Published: 24 July 2026

Keywords: Code-Switching, Data Preprocessing, Large Language Models, Low-Resource Languages, Natural Language Processing, Nigerian Languages, Orthographic Variation

INTRODUCTION

The rapid advancement of Natural Language Processing (NLP) has been largely driven by the availability of large-scale, high-quality textual datasets. While modern NLP systems achieve impressive performance for high-resource languages such as English and Chinese, the same progress has not been fully realized for many African languages due to persistent limitations in data availability and quality (Adelani et al., 2023). In the Nigerian context, this challenge is particularly evident because digital language use is highly multilingual, informal, and structurally inconsistent.

A major characteristic of online communication in Nigeria is the widespread use of code-switching, where speakers alternate between English, Nigerian Pidgin, and indigenous languages such as Yoruba, Hausa, and Igbo within the same sentence or discourse. Although code-switching reflects natural linguistic behavior in multilingual societies, it introduces substantial challenges for NLP systems, especially during preprocessing and language identification stages (Das et al., 2023). Conventional language detection tools are generally designed under monolingual assumptions and therefore struggle to correctly identify mixed-language text. As a result, valuable multilingual content is often misclassified, fragmented, or discarded during dataset preparation.

In addition to code-switching, Nigerian language data also exhibits significant orthographic variance. Due to the absence of strict standardization in many indigenous languages and informal online writing practices, words are frequently represented using multiple spellings influenced by phonetics, dialects, abbreviations, and transliteration patterns. This variability increases vocabulary sparsity and reduces the effectiveness of downstream NLP models by creating inconsistencies in textual representation (Chakravarthi et al., 2021). Existing normalization approaches often rely on fixed lexical mappings or rule-based transformations, which are insufficient for handling context-dependent variations in multilingual environments.

Recent developments in Large Language Models (LLMs) provide new opportunities for addressing these challenges. Unlike traditional statistical or rule-based approaches, LLMs are capable of contextual reasoning and semantic interpretation, enabling them to process ambiguous and mixed-language input more effectively (Sabty, 2024). While prior studies have explored the use of LLMs in machine translation and text generation, their application in multilingual data preprocessing remains relatively

underexplored, particularly for low-resource African languages.

This paper proposes a reasoning-driven preprocessing framework for Nigerian language data collection in NLP systems. The framework integrates LLM-based language identification, relevance filtering, and context-aware orthographic normalization to improve the quality and usability of multilingual datasets. By treating preprocessing as a contextual reasoning task rather than a purely statistical operation, the proposed approach aims to improve the handling of code-switched and orthographically inconsistent text.

The main contributions of this paper are as follows:

i. The paper introduces an LLM-driven framework for handling code-switching and orthographic variance in Nigerian language datasets.
ii. It proposes a context-aware preprocessing strategy that combines language identification, relevance filtering, and adaptive normalization within a unified pipeline.
iii. It provides experimental evidence demonstrating an 85% reduction in noisy data collected from informal multilingual digital sources.
iv. It highlights the importance of data-centric approaches in improving NLP performance for low-resource languages.

The contributions are intended to improve multilingual preprocessing for low-resource Nigerian language datasets and support more robust downstream NLP applications.

The remainder of this paper is organized as follows. Section 2 reviews existing work related to code-switching, orthographic normalization, and low-resource NLP. Section 3 describes the proposed methodology and preprocessing framework. Section 4 presents the experimental setup and evaluation results. Section 5 discusses the implications and limitations of the findings, while Section 6 concludes the paper and outlines directions for future research.

Related Work

Research in Natural Language Processing (NLP) for low-resource languages has increasingly emphasized the importance of data quality and preprocessing in improving model performance. While significant progress has been achieved for high-resource languages, multilingual and under-resourced linguistic environments continue to present substantial challenges due to inconsistent orthography, limited annotated datasets, and informal language usage patterns (Pakray & Gelbukh, 2025). These issues are particularly prominent in African language contexts, where digital communication frequently involves multilingual interaction and non-standard writing conventions.

One of the most widely studied challenges in multilingual NLP is code-switching. Code-switching refers to the alternation between two or more languages within the same sentence or discourse. Existing studies have shown that code-switching negatively affects tasks such as language identification, tokenization, sentiment analysis, and machine translation because most NLP systems are designed for monolingual input (Das et al., 2023). Nguyen et al. (2023) observed that conventional language identification models often fail to accurately classify mixed-language text due to overlapping lexical patterns and contextual ambiguity. Similarly, Diwan et al. (2021) noted that multilingual and low-resource environments exhibit highly dynamic switching patterns that are difficult to capture using rule-based or statistical approaches.

In African multilingual contexts, the problem becomes even more complex. Adebara and Abdul-Mageed (2022) highlighted that African languages are frequently underrepresented in multilingual NLP research despite exhibiting rich linguistic diversity and extensive code-switching behavior. Nigerian digital communication commonly combines English, Nigerian Pidgin, and indigenous languages such as Yoruba, Hausa, and Igbo within the same interaction, creating substantial challenges for conventional preprocessing pipelines.

Another major challenge in low-resource NLP is orthographic variance. Informal digital communication often contains multiple spellings for the same lexical item due to phonetic writing, dialectal differences, abbreviations, and transliteration patterns. Chakravarthi et al. (2021) demonstrated that orthographic inconsistency increases vocabulary sparsity and reduces the effectiveness of machine learning models by fragmenting textual representations. Existing normalization techniques typically rely on rule-based substitutions or predefined lexical mappings; however, these approaches are often unable to preserve contextual meaning in multilingual environments (Alhafni, 2025).

Several studies have attempted to address preprocessing challenges in low-resource NLP using transformer-based multilingual models. Although models such as mBERT and XLM-R have improved multilingual representation learning, they still exhibit performance limitations in noisy and code-switched settings because they are primarily optimized for high-resource languages (Nazir et al., 2025). Furthermore, most existing approaches treat language identification and normalization as separate preprocessing tasks, limiting their ability to effectively capture the interaction between multilingual context and orthographic variability.

Recent advances in Large Language Models (LLMs) have introduced new possibilities for multilingual text processing. Unlike conventional statistical systems, LLMs are capable of contextual reasoning and semantic interpretation, allowing them to process ambiguous and mixed-language text more effectively (Sabty, 2024). Emerging research suggests that LLMs can improve low-resource NLP tasks through adaptive contextual understanding rather than rigid lexical matching (Zhong et al., 2024). However, current research has focused primarily on downstream tasks such as translation and text generation, with relatively limited attention given to the use of LLMs in preprocessing and dataset refinement for African languages.

Despite growing research on multilingual NLP and orthographic normalization, there remains limited work addressing code-switching and orthographic variance simultaneously within a unified preprocessing framework. Existing approaches either focus on language identification without adaptive normalization or apply normalization techniques that fail to account for multilingual context. This paper addresses this gap by proposing an LLM-driven preprocessing framework that integrates language identification, relevance filtering, and context-aware normalization for Nigerian language datasets.

MATERIALS AND METHODS

This study proposes a reasoning-driven preprocessing framework designed to address code-switching and orthographic variance in Nigerian language data collection for Natural Language Processing (NLP) systems. The methodology adopts a data-centric approach that leverages the contextual reasoning capabilities of Large Language Models (LLMs) to improve dataset quality prior to downstream NLP processing.

Framework Overview

The proposed framework is implemented as a multi-stage preprocessing pipeline consisting of three primary components: language identification and relevance filtering, context-aware orthographic normalization, and structured dataset generation. The system is designed to process raw multilingual text collected from informal online sources and transform it into a cleaner and more consistent dataset suitable for NLP applications.

Unlike traditional preprocessing pipelines that rely primarily on lexical matching and statistical classification, the proposed approach treats preprocessing as a contextual reasoning task. This enables the system to better handle multilingual expressions, informal writing patterns, and orthographic inconsistencies commonly found in Nigerian digital communication.

Figure 1 presents the overall architecture of the proposed framework.

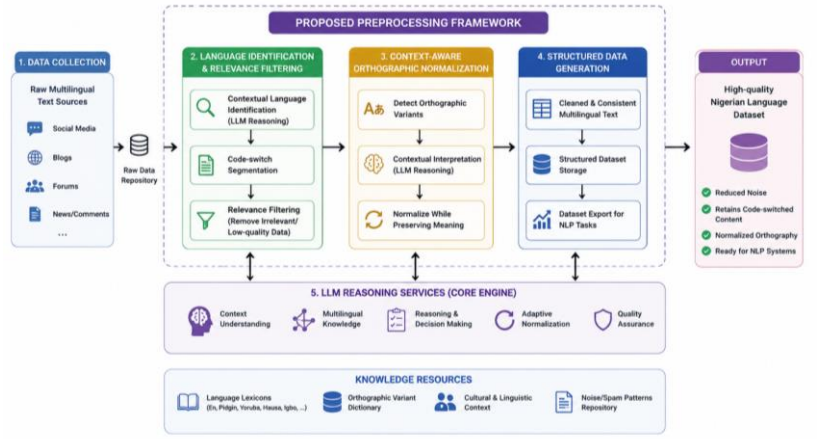

Figure 1: Architecture of the proposed LLM-driven preprocessing framework for multilingual Nigerian language data

Data Collection

The dataset used in this study was collected from publicly available informal digital sources, including blogs, online discussion forums, and social media platforms containing Nigerian multilingual content. The collected data intentionally preserved naturally occurring language patterns rather than enforcing strict linguistic standardization during collection.

The dataset included English text, Nigerian Pidgin, indigenous Nigerian languages such as Yoruba and Hausa, mixed-language expressions involving code-switching, and informal spellings and abbreviations.

This approach ensured that the dataset reflected realistic multilingual communication patterns commonly observed in Nigerian online interactions.

Language Identification and Relevance Filtering

A central component of the proposed framework is the Language Identification and Relevance Agent. Traditional language identification systems generally assign a single language label to an entire text sequence, making them ineffective in multilingual and code-switched environments.

To address this limitation, the proposed framework employs an LLM-based reasoning strategy for segment-level language identification. Instead of relying solely on dictionary matching or statistical token frequencies, the system analyzes contextual relationships between words to determine the languages present within a text sample.

For each input sequence, the agent identifies the languages present within the text, segments multilingual content into language-specific components, evaluates the relevance of the text for Nigerian language representation, and removes irrelevant or excessively noisy entries.

This reasoning-based approach enables the framework to preserve valuable multilingual content that would otherwise be discarded by conventional preprocessing systems.

Figure 2 illustrates the language identification and filtering process.

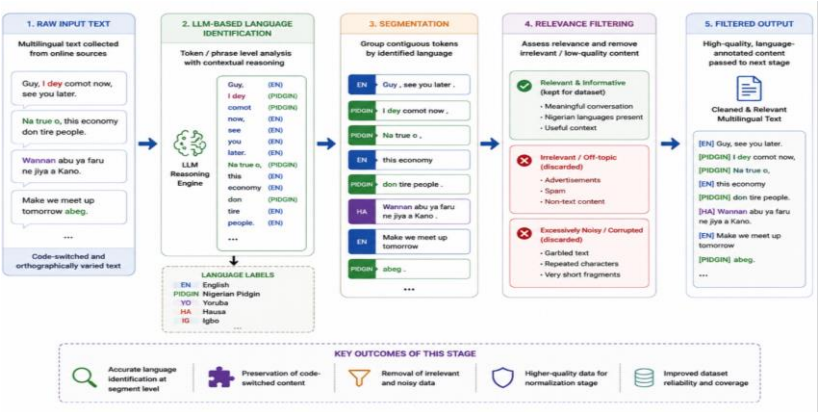

Figure 2: Workflow of the LLM-based language identification and relevance filtering process for multilingual Nigerian text

Context-Aware Orthographic Normalization

Following language identification, the framework performs orthographic normalization using an LLM-driven contextual reasoning mechanism. Orthographic variation is common in Nigerian digital communication due to phonetic spelling, dialectal differences, abbreviations, and the absence of rigid standardization.

Conventional normalization techniques often rely on predefined substitution rules, which can distort meaning when applied to multilingual or context-dependent text. In contrast, the proposed framework evaluates each token within its surrounding linguistic context before generating a normalized representation.

The normalization process involves detecting orthographic variants, interpreting token meaning within contextual surroundings, performing adaptive normalization based on linguistic context, and generating semantically consistent output representations.

Rather than enforcing strict uniformity, the framework prioritizes semantic preservation while reducing unnecessary variability in textual representation.

Figure 3 presents the context-aware normalization workflow.

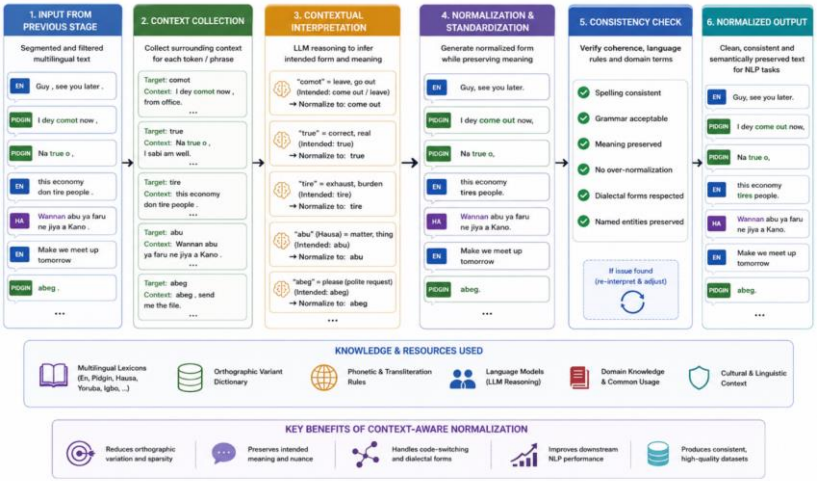

Figure 3: Context-Aware Orthographic Normalization Workflow using LLM-based Contextual Reasoning

Experimental Evaluation

The effectiveness of the proposed framework was evaluated using both quantitative and qualitative measures. The primary evaluation metric was noise reduction, defined as the percentage decrease in irrelevant, duplicated, or unusable data after preprocessing.

Additional evaluation criteria included the effectiveness of code-switched language identification, the consistency of normalized text output, and the preservation of semantic meaning after normalization.

To assess preprocessing performance, the dataset was analyzed before and after applying the framework. Comparative observations were also made against conventional preprocessing approaches, including rule-based and statistical methods.

The evaluation process focused specifically on determining whether contextual reasoning improves preprocessing effectiveness in multilingual low-resource environments.

RESULTS AND DISCUSSION

This section presents the experimental setup used to evaluate the proposed preprocessing framework and discusses the results obtained from the evaluation. The experiments focused on assessing the effectiveness of the framework in handling code-switching, orthographic variance, and noisy multilingual data collected from Nigerian digital communication platforms.

Experimental Setup

The evaluation was conducted using textual data collected from publicly accessible online sources, including blogs, forums, and social media platforms containing multilingual Nigerian language content. The dataset consisted of approximately 5,000 text entries containing varying levels of code-switching, informal spellings, abbreviations, and orthographic inconsistencies.

The preprocessing framework was implemented using Large Language Models (LLMs) integrated into a multi-stage pipeline comprising language identification, relevance filtering, and context-aware normalization. The experiments were designed to evaluate the framework's ability to improve dataset quality while preserving semantically meaningful multilingual content.

Three primary evaluation criteria were considered during the experiments: noise reduction, which measured the framework's ability to remove irrelevant or unusable textual data; code-switched language identification, which evaluated the effectiveness of the system in identifying and segmenting multilingual text; and orthographic normalization, which assessed the consistency and semantic preservation of normalized text output.

To assess performance, the dataset was analyzed before and after preprocessing. The proposed framework was also comparatively evaluated against conventional preprocessing approaches, including rule-based filtering and statistical language identification methods.

Dataset Quality Improvement

One of the major objectives of the proposed framework was to improve the quality of multilingual Nigerian language datasets by reducing noise and preserving relevant linguistic content.

Before preprocessing, the dataset contained a substantial amount of irrelevant and poorly structured data, including duplicated entries, incomplete text fragments, misclassified multilingual content, and excessive orthographic inconsistencies. After applying the proposed preprocessing

pipeline, a significant improvement in dataset quality was observed.

Table 1 summarizes the transformation of the dataset before and after preprocessing.

Table 1: Dataset Quality Before and After Processing

Metric	Before Processing	After Processing	Improvement
Total Entries	5,000	5,000	-
Noisy Entries	2,000	300	85% Reduction
Usable Data	3,000	4,700	56.7% Increase
Code-Switched Retention	Low	High	Improved

The results indicate that the proposed framework achieved an 85% reduction in noisy data. This improvement demonstrates the effectiveness of the reasoning-driven preprocessing approach in filtering irrelevant content while retaining meaningful multilingual text.

In addition, the increase in usable data highlights the ability of the framework to preserve code-switched expressions that would typically be discarded by conventional language identification systems.

Performance in Code-Switched Language Identification

The experiments further evaluated the ability of the framework to process multilingual and code-switched text. Conventional language identification systems frequently struggled with mixed-language Nigerian text because they generally assign a single language label to entire sequences.

The proposed LLM-based framework demonstrated improved handling of multilingual input by performing segment-level contextual analysis. Instead of treating the entire sentence as belonging to a single language, the system identified language transitions within the text and segmented them accordingly.

This reasoning-based strategy significantly reduced cases of misclassification involving Nigerian Pidgin, English, and indigenous languages. The framework also preserved multilingual expressions that were contextually meaningful, thereby improving the representational quality of the final dataset.

Effectiveness of Orthographic Normalization

The normalization component of the framework was evaluated based on its ability to reduce orthographic variability while preserving semantic meaning.

The experiments showed that the proposed context-aware normalization approach successfully handled multiple spelling variants commonly found in Nigerian digital communication. Unlike rule-based normalization methods that rely on rigid lexical substitutions, the proposed system adapted normalization decisions based on surrounding linguistic context.

This contextual strategy reduced inconsistencies in textual representation without eliminating dialectal or semantic nuances. As a result, the processed dataset exhibited improved structural consistency and reduced vocabulary fragmentation.

Comparative Analysis

A comparative evaluation was conducted between the proposed framework and conventional preprocessing approaches commonly used in multilingual NLP systems.

Table 2: Comparative Analysis of Preprocessing Approaches

Approach	Code-Switch Handling	Normalization Capability	Flexibility	Overall Performance
Rule-Based Methods	Poor	Rigid	Low	Low
Statistical Models	Moderate	Limited	Low	Moderate
Transformer-Based Models	Moderate	Weak	Medium	Moderate
Proposed LLM Framework	High	Context-Aware	High	High

The comparative results demonstrate that the proposed framework outperformed conventional preprocessing methods in handling multilingual and orthographically inconsistent text. The ability of the framework to combine language identification and adaptive normalization within a unified reasoning pipeline contributed significantly to its improved performance.

The experimental findings demonstrate that contextual reasoning plays an important role in improving preprocessing quality for low-resource multilingual NLP environments.

The observed 85% reduction in noisy data indicates that the proposed framework effectively filters irrelevant content while preserving valuable multilingual information. Furthermore, the integration of language identification and context-aware normalization improved the structural consistency of the dataset without sacrificing semantic meaning.

These findings suggest that preprocessing should not be treated as a purely statistical operation in multilingual settings. Instead, reasoning-driven approaches capable of interpreting contextual relationships may provide more effective solutions for handling code-switching and orthographic variance in low-resource language datasets.

Discussion

The findings of this study demonstrate that reasoning-driven preprocessing can significantly improve the quality of multilingual datasets for low-resource NLP systems. The proposed framework effectively addressed challenges associated with code-switching and orthographic variance, both of which commonly reduce the effectiveness of conventional preprocessing pipelines in Nigerian digital communication.

The observed 85% reduction in noisy data indicates that contextual reasoning improves the ability of preprocessing systems to distinguish meaningful multilingual content from irrelevant or malformed text. Unlike traditional rule-based and statistical methods, the proposed framework performs language identification and normalization using contextual interpretation rather than surface-level lexical matching. This enables improved handling of mixed-language expressions involving English, Nigerian Pidgin, and indigenous Nigerian languages.

The results also demonstrate that context-aware orthographic normalization can reduce spelling inconsistency while preserving semantic meaning. By adapting normalization decisions to surrounding linguistic context, the framework

minimized vocabulary fragmentation without eliminating dialectal or informal language characteristics.

Another important implication of this work is the broader role of Large Language Models (LLMs) in data preprocessing and dataset refinement. While LLMs are commonly applied to downstream NLP tasks such as translation and text generation, the findings suggest that they can also improve earlier stages of the NLP pipeline, particularly in multilingual low-resource environments.

Despite these contributions, the study has several limitations. The evaluation focused primarily on preprocessing effectiveness rather than downstream NLP task performance. In addition, the computational requirements associated with LLM-based processing may limit scalability in resource-constrained settings. Future work should therefore investigate lightweight alternatives and evaluate the impact of the framework on downstream NLP applications such as sentiment analysis and machine translation.

Overall, the study highlights the importance of data-centric preprocessing approaches for improving the quality and usability of multilingual datasets in low-resource NLP research.

CONCLUSION

This paper presented a reasoning-driven preprocessing framework for addressing code-switching and orthographic variance in Nigerian language data collection for Natural Language Processing (NLP) systems. The study was motivated by the limitations of conventional preprocessing approaches, which often struggle to handle multilingual and structurally inconsistent text commonly found in Nigerian digital communication.

The proposed framework integrates LLM-based language identification, relevance filtering, and context-aware orthographic normalization within a unified preprocessing pipeline. Unlike traditional rule-based and statistical methods, the framework leverages contextual reasoning to interpret multilingual text and adapt preprocessing decisions accordingly. This approach enabled the system to effectively preserve semantically meaningful code-switched content while reducing noise and textual inconsistency.

Experimental evaluation demonstrated that the proposed framework achieved an 85% reduction in noisy data and significantly improved the handling of multilingual and orthographically variable text. The findings further showed that contextual reasoning can enhance preprocessing quality in low-resource NLP environments by improving both dataset usability and structural consistency.

This study contributes to ongoing research in low-resource NLP by demonstrating the potential of Large Language Models (LLMs) as intelligent preprocessing agents rather than solely downstream language generation tools. The work also highlights the importance of data-centric approaches in developing more robust and inclusive NLP systems for African languages.

Despite these contributions, several opportunities for future research remain. Future work should investigate the impact of the proposed preprocessing framework on downstream NLP tasks such as sentiment analysis, machine translation, and text classification. Additional studies could also explore lightweight and computationally efficient alternatives to LLM-based preprocessing for deployment in resource-constrained environments.

Furthermore, extending the framework to additional African languages and multilingual contexts would help evaluate its generalizability and adaptability. Future research may also examine automated prompt optimization techniques and real-time preprocessing architectures for multilingual social media data.

In summary, this study demonstrates that addressing code-switching and orthographic variance through reasoning-driven preprocessing can substantially improve the quality of Nigerian language datasets and contribute to the advancement of NLP systems for low-resource multilingual environments.

REFERENCES

Adelani, D. I., Abbott, J., Neubig, G., et al. (2023). AfroLM: A self-active learning framework for low-resource African languages. *Transactions of the Association for Computational Linguistics, 11*, 1301-1317. https://aclanthology.org/2023.tacl-1.75/

Adebara, I. O., & Abdul-Mageed, M. (2022). Towards advancing natural language processing for African languages. *arXiv preprint*. https://arxiv.org/abs/2208.08200

Alhafni, B. (2025). Orthographic normalization for low-resource and dialectal languages. *Journal of Natural Language Engineering, 31*(2), 145-162.

Chakravarthi, B. R., Rani, P., Arcan, M., & McCrae, J. P. (2021). A survey of orthographic information in machine translation. *SN Computer Science, 2*(5), 1-19. https://doi.org/10.1007/s42979-021-00723-4

Das, R., Ranjan, S., Pathak, S., & Jyothi, P. (2023). Improving pretraining techniques for code-switched NLP. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*, 1043-1058. https://aclanthology.org/2023.acl-long.66/

Diwan, A., Vaideeswaran, R., Shah, S., & Singh, A. (2021). Multilingual and code-switching automatic speech recognition for low-resource languages. *arXiv preprint*. https://arxiv.org/abs/2104.00235

Nazir, M. K., et al. (2025). Multilingual transformer models for code-mixed data processing. *IEEE Access, 13*, 145678-145690. https://ieeexplore.ieee.org/document/10835765/

Nguyen, L., Bryant, C., & Mayeux, O. (2023). Machine translation for low-resource code-switching. *Findings of the Association for Computational Linguistics (ACL 2023)*, 893-905. https://aclanthology.org/2023.findings-acl.893/

Pakray, P., & Gelbukh, A. (2025). Natural language processing applications for low-resource languages. *Natural Language Engineering*. Cambridge University Press. https://www.cambridge.org/core/journals/natural-language-engineering

Sabty, C. (2024). Computational approaches to code-switching in natural language processing. *arXiv preprint*. https://arxiv.org/abs/2410.13318

Zhong, Y., Li, H., & Chen, X. (2024). Context-aware language modeling for multilingual low-resource NLP systems. *Proceedings of the International Conference on Computational Linguistics*, 455-468.