



(RESEARCH ARTICLE)



Hybrid neural model for Hausa text auto completion: addressing data scarcity for a low-resource language

Abubakar Safiyanu *, Buhari Aliyu, Hadiza Ibrahim Aminu and Aliyu Abdullahi

Department Of Computer Engineering Technology, Jigawa State Polytechnic Dutse, Duste, Nigeria.

World Journal of Advanced Research and Reviews, 2026, 30(02), 379-387

Publication history: Received on 26 March 2026; revised on 02 May 2026; accepted on 05 May 2026

Article DOI: <https://doi.org/10.30574/wjarr.2026.30.2.1199>

Abstract

The Hausa language, a major linguistic vehicle for over 70 million people in West Africa, remains critically underserved by contemporary Natural Language Processing (NLP) technologies, exacerbating the digital divide. This paper presents the design and evaluation of a dedicated neural text autocomplete system to address this gap. Confronting the fundamental challenge of data scarcity, we developed a hybrid corpus of 50,000 sentences, merging authentic Hausa text with algorithmically generated sentences created via a rule-based generator. For prediction, we implemented a stacked Long Short-Term Memory (LSTM) neural network, enhanced with a large language model (LLM) fallback mechanism for robustness. The system achieved a Top-1 accuracy of 92.4% and a Top-5 accuracy of 98.6% on a held-out test set, significantly outperforming traditional trigram (Top-3: 82.5%) and simple RNN (Top-3: 89.1%) baselines. A user study with 30 Hausa speakers confirmed its practical utility, demonstrating a 35% average increase in typing speed and a 91% acceptance rate for the primary suggestion. This work provides a reproducible framework for developing NLP tools in low-resource settings, introduces a novel hybrid dataset for Hausa which we publicly release, and delivers empirical evidence of tangible benefits for users. Our findings offer a scalable blueprint for enhancing digital inclusivity for underserved language communities.

Keywords: Hausa language; Text autocompletion; Low-resource NLP; LSTM; Neural text prediction; Digital inclusion; African language technology; Hybrid dataset; Natural Language Processing; Socio-technical systems

1. Introduction

The proliferation of intelligent text autocompletion has become a hallmark of modern digital communication, significantly enhancing typing speed, accuracy, and user experience. Powered by advances in deep learning, particularly transformer-based architectures, these systems have reached high levels of sophistication for globally dominant languages such as English and Mandarin [1, 2]. However, this technological progress has been highly asymmetrical, creating a growing digital language divide. Hundreds of millions of speakers of low-resource languages those with limited digitally available text for training machine learning models remain excluded from the benefits of such assistive technologies [3].

Hausa, a Chadic language spoken natively by over 50 million people and used as a lingua franca by millions more across West Africa, epitomizes this challenge [4]. As a language of major media, education, commerce, and social interaction in the region, the absence of robust digital tools stifles productivity and inclusivity. While major platforms like Google Gboard or Microsoft SwiftKey offer basic support for Hausa, their predictive models primarily optimized for high-resource languages often generate inaccurate or contextually irrelevant suggestions. This inadequacy stems from a lack of large, curated, and linguistically representative training data, which is the cornerstone of modern data-driven NLP [5].

* Corresponding author: Abubakar Safiyanu

The development of effective language technology for Hausa is further complicated by specific linguistic and sociolinguistic factors. These include dialectal variations (e.g., between Kano, Sokoto, and Katsina dialects), frequent code-switching with English and Arabic in both written and spoken forms, and orthographic nuances [6, 7]. Consequently, building a functional autocomplete system requires more than mere linguistic translation of existing models; it necessitates an approach tailored to the data-scarce environment and the unique characteristics of the language.

In this research, we address this critical gap by presenting a dedicated neural text autocomplete system for the Hausa language. Our work is guided by the following research question: How can an effective, context-aware autocomplete system be developed for a low-resource language like Hausa despite severe training data constraints? To answer this, we propose and validate a novel methodology based on two key strategies: (1) the creation of a hybrid text corpus that strategically combines authentic and synthetically generated data to overcome scarcity, and (2) the implementation of a stacked Long Short-Term Memory (LSTM) neural network, augmented with a large language model (LLM) fallback mechanism, chosen for its efficiency in learning sequential patterns from limited data.

The primary contributions of this research are fourfold:

- **A Novel Hybrid Dataset:** We compile and publicly release a curated dataset of 50,000 Hausa sentences, created through a hybrid method designed to maximize linguistic coverage and grammatical validity within a low-resource context.
- **A High-Performance Predictive Model:** We design, train, and optimize a stacked LSTM model that achieves state-of-the-art predictive performance for Hausa, demonstrating the viability of this architecture for morphologically rich, low-resource languages.
- **A Comprehensive Evaluation Framework:** We evaluate our system using both standard computational metrics (e.g., Top-K accuracy, F1-score) and, crucially, a practical user study with 30 participants to assess real-world usability, typing efficiency, and acceptance.
- **A Reproducible Framework:** We provide a complete blueprint for developing similar assistive technologies for other low-resource languages, thereby contributing to the broader goal of global digital inclusivity.

The remainder of this paper is structured as follows. Section 2 reviews related work in autocomplete systems and NLP for low-resource and African languages. Section 3 details our methodology, including data curation, model architecture, and training procedures. Section 4 presents the experimental results and analysis. Section 5 discusses the implications of our findings, acknowledges limitations, and suggests directions for future work. Finally, Section 6 concludes the paper.

2. Related Work

2.1. Evolution of Text Prediction Technologies

Early autocomplete systems were primarily dictionary-based, relying on static word lists to suggest completions for typed prefixes. While these systems reduced keystrokes, they lacked contextual awareness and failed to adapt to user writing patterns [8]. The introduction of predictive text input on mobile devices, such as the T9 system, represented a step forward by mapping multiple letters to numeric keys and predicting words from a built-in dictionary. However, its static vocabulary and inability to learn from usage were major limitations [9]. The adoption of statistical language models, notably n-grams, marked a shift towards context-aware prediction by calculating the probability of a word given its preceding n-* words. While an improvement, n-gram models are hindered by data sparsity and an inability to capture long-range dependencies in text [10]. The breakthrough for modern autocomplete systems came with the rise of neural network architectures capable of learning distributed representations of language.

2.2. Machine Learning for Autocomplete Systems

Machine learning, particularly deep learning, has become the cornerstone of state-of-the-art autocomplete systems. The development of word embedding techniques like Word2Vec and GloVe enabled models to represent words in a continuous vector space, capturing semantic and syntactic relationships [11, 12]. For sequence modeling, Recurrent Neural Networks (RNNs) and their variant, Long Short-Term Memory (LSTM) networks, became prominent due to their ability to process sequential data and retain context over longer text spans, effectively addressing the vanishing gradient problem of traditional RNNs [13, 14]. The current paradigm is dominated by transformer-based models, which utilize self-attention mechanisms to weigh the importance of all words in a sequence simultaneously, enabling superior context modeling. Models like GPT and BERT have set new benchmarks for language understanding and generation tasks, including text prediction [15, 16, 17].

2.3. Autocomplete Systems for Low-Resource Languages

Despite these advancements, the focus of research and commercial development remains skewed toward high-resource languages. Commercial systems like Google's Gboard, Smart Compose, and Microsoft's SwiftKey are optimized primarily for languages like English, French, and Chinese [18]. For low-resource languages, the scarcity of large, digital text corpora poses a fundamental challenge for training data-intensive models like transformers [19]. Recent efforts in multilingual NLP have led to models like mBERT and XLM-R, which leverage cross-lingual transfer learning, offering a potential pathway for supporting underserved languages [19]. In the African context, initiatives are emerging to build resources and benchmarks, as noted in surveys of Nigerian NLP [20] and through the creation of sentiment analysis datasets for languages like Hausa [21, 22]. However, applied research focused on building end-user tools, such as reliable predictive text, remains scarce.

2.4. The Hausa NLP Context and Identified Gap

Hausa, a major West African language, exemplifies the challenges of low-resource language NLP. While foundational linguistic resources exist, computational research has only recently gained momentum. Systematic reviews of Hausa NLP highlight progress in areas like sentiment analysis and machine translation but also underscore a persistent lack of tools for interactive, generative tasks like text autocompletion [23]. Specific challenges include dialectal variation, frequent code-switching with English and Arabic, and orthographic inconsistencies, which are not accounted for in generic predictive systems [24, 25]. Although recent work has begun to address related issues like anomaly correction in Hausa text [26], a dedicated, data-driven autocomplete system built and evaluated for Hausa has not been previously developed. This study directly addresses this gap by constructing a tailored Hausa dataset, designing a context-sensitive LSTM-based model, and evaluating its performance against both computational metrics and real-world usability.

3. Methodology

This research employs a quantitative experimental design to develop and evaluate a hybrid neural system for Hausa text autocompletion. The methodology is structured around four pillars:

- The creation of a hybrid dataset to overcome resource scarcity,
- The design and training of a stacked LSTM model,
- The integration of a large language model (LLM) for fallback prediction,

A multi-modal evaluation strategy.

3.1. Data Collection and Preparation

Addressing the absence of large-scale Hausa text corpora was the primary challenge. We developed a two-pronged data strategy.

3.1.1. Hybrid Dataset Creation

A custom HausaDataGenerator class was implemented to systematically create training data. This involved:

- Vocabulary Curation: Expanding lists of common Hausa nouns, verbs, adjectives, pronouns, conjunctions, prepositions, and question words.
- Synthetic Sentence Generation: Using defined syntactic templates (e.g., "{pronoun} {verb} {noun}", "Mu {verb} {noun} tare") to generate grammatically valid sentences. This ensured coverage of diverse sentence structures.
- Integration of Authentic Text: A curated set of common, authentic Hausa phrases was incorporated to improve naturalness and model generalization.

The final dataset comprised 50,000 sentences, split into 40,000 for training, 7,000 for validation, and 3,000 for testing.

3.1.2. Preprocessing and Sequence Preparation

For model training, sentences were transformed into predictor-target pairs. Each sentence was converted into a series of n-gram sequences where the first n words predict the n+1 word (e.g., for "Mu je kasuwa yau," one sequence is: Input: ["Mu", "je", "kasuwa"] → Target: "yau"). Sequences were tokenized using the Keras Tokenizer (vocabulary size: 3,000), padded to a fixed length, and the target word was one-hot encoded.

3.2. System Architecture and Model Design

The core autocomplete system is a hybrid architecture designed for accuracy and robustness in a low-resource context.

3.2.1. Primary Predictor: Stacked LSTM Model

A stacked Long Short-Term Memory (LSTM) network was selected as the primary predictive engine due to its proven capability in modeling sequential dependencies in language [13, 14]. The model was configured as follows:

- Embedding Layer: 100-dimensional dense representations.
- LSTM Layers: Two stacked layers (150 units in the first layer with `return_sequences=True`, 100 units in the second).
- Regularization: A dropout layer (`rate=0.2`) to mitigate overfitting.
- Output Layer: A dense layer with softmax activation over the 3,000-word vocabulary.

3.2.2. Fallback and Enhancement: Large Language Model (LLM) Integration

To handle out-of-vocabulary terms and complex, long-range contextual cues, the system integrates a pre-trained multilingual LLM as a fallback predictor. When the LSTM model's confidence score for its top prediction is below a defined threshold, the LLM is queried to generate a contextually appropriate suggestion. This hybrid LSTM+LLM design ensures high reliability.

3.3. Training and Optimization Procedures

The LSTM model was trained using sparse categorical cross-entropy loss and the Adam optimizer (initial learning rate: 0.001) [27]. Key hyperparameters included a batch size of 64, training for 100 epochs, and a sequence length of 5. Strategies like early stopping (patience: 10 epochs) and learning rate reduction on plateau were employed to optimize convergence and prevent overfitting.

3.4. Evaluation Framework

System performance was assessed through computational metrics and real-world usability testing.

- Predictive Performance Metrics: We evaluated Top-K accuracy ($K=1, 3, 5, 10$), precision, recall, and F1-score on the test set.
- Comparative Baseline Models: Performance was benchmarked against a statistical Trigram model and a simple RNN to contextualize improvements.
- User-Centric Evaluation: A controlled study with 30 Hausa-speaking participants measured typing speed improvement and prediction acceptance rate to assess practical utility.

This detailed methodology ensures the system is built on a reproducible, technically sound foundation and is evaluated against both algorithmic and human-centric standards.

4. Results & Analysis

4.1. Dataset Characteristics

The final hybrid corpus contained 50,000 sentences with 12,500 unique tokens. Table 1 summarizes the distribution of primary word types, illustrating the dataset's linguistic coverage, which is crucial for training a robust language model.

Table 1 Vocabulary Distribution in the Hausa Corpus

Word Type	Count	Percentage (%)
Nouns	5,100	40.8
Verbs	3,500	28.0
Adjectives	1,800	14.4
Adverbs	1,200	9.6
Function Words	900	7.2

4.2. Model Performance Metrics

The stacked LSTM model demonstrated strong predictive accuracy on the held-out test set of 3,000 sentences. Performance was measured using Top-K accuracy, which indicates whether the correct next word appears within the top K suggestions a critical metric for user-facing autocomplete systems.

Table 2 Top-K Predictive Accuracy of the LSTM Model

Top-K	Accuracy (%)
1	92.4
3	97.1
5	98.6
10	99.3

The high Top-5 accuracy (98.6%) indicates that the correct word was virtually always present within the first five suggestions, providing a highly usable experience. Standard classification metrics further confirmed model robustness, with a precision of 95.2%, recall of 94.7%, and an F1-score of 94.95%.

4.3. Comparative Analysis with Baseline Models

To contextualize the performance of the proposed LSTM model, it was compared against two common baseline approaches: a statistical Trigram model and a simple Recurrent Neural Network (RNN). The results, summarized in Table 3, show a clear and significant advantage for the stacked LSTM architecture.

Table 3 Comparative Performance of Models

Model	Top-1 Accuracy (%)	Top-3 Accuracy (%)	Top-5 Accuracy (%)	F1-Score (%)
Trigram(3-gram)	75.2	82.5	85.1	80.2
Simple RNN	85.3	89.1	91.2	88.0
Proposed LSTM	92.4	97.1	98.6	94.95

The LSTM model's superior performance is attributed to its ability to capture longer-range contextual dependencies and manage the vanishing gradient problem more effectively than the simpler baselines [13].

4.4. User Study: Practical Usability and Efficiency

Beyond computational metrics, a user study with 30 proficient Hausa speakers was conducted to evaluate real-world utility. Participants performed standardized typing tasks with and without the autocomplete system activated. Key findings include:

- **Typing Speed:** Average typing speed increased by 34.3%, from 35 Words Per Minute (WPM) to 47 WPM.
- **Suggestion Acceptance:** The Top-1 suggestion provided by the system was accepted by users 91% of the time.
- **Error Reduction:** Participants reported a subjective decrease in typographical errors and typing fatigue.

These results validate the system's practical effectiveness in enhancing digital communication efficiency for Hausa speakers.

4.5. Error Analysis and Limitations

A detailed analysis of incorrect predictions revealed consistent patterns, informing areas for future improvement. Table 4 illustrates the relationship between word frequency and prediction accuracy, highlighting one key limitation.

Table 4 Word Frequency vs. Prediction Accuracy

Frequency Category	Sample Words	Accuracy (%)
High (>500)	gida, makaranta	99.2
Medium (100-500)	kasuwa, tafiya	95.5
Low (<100)	garin, gidan gona	87.4

The strong positive correlation between word frequency and accuracy indicates that rare words are more challenging for the model. Other error patterns include:

- **Morphological Complexity:** Errors occasionally occurred with morphologically rich words (e.g., verbs with tense/aspect affixes) where the model failed to generalize from root forms.
- **Long-Range Context:** A slight degradation in Top-1 accuracy was observed for sentences exceeding 15 words, indicating a limitation of the LSTM's sequential memory in capturing very distant dependencies.

4.6. Summary of the Findings

The results confirm that the hybrid LSTM+LLM autocomplete system successfully addresses the core challenge of data scarcity for Hausa. It achieves:

- **High Predictive Accuracy:** Exceeding 92% Top-1 and 98% Top-5 accuracy.
- **Superiority Over Baselines:** Outperforming traditional n-gram and simple RNN models by a significant margin (Table 3).
- **Tangible User Benefits:** Delivering measurable improvements in typing speed (34.3% increase) and user acceptance (91% acceptance rate).
- **Identified Improvement Paths:** Clear error patterns (Table 4) point towards future work in data augmentation and architectural enhancements for handling rare words and long contexts.

4.7. Interpretation of the Findings

The system's high predictive accuracy (Top-5: 98.6%) and its substantial positive impact on user typing speed (+34.3%) and acceptance (91%) confirm the core hypothesis: that a tailored, data-efficient neural approach can effectively bridge the technological gap for Hausa speakers.

The superior performance of the LSTM model over statistical (Trigram) and simpler neural (RNN) baselines (Table 3) underscores the importance of architecture choice. LSTMs, with their gated memory, are particularly well-suited for morphologically rich languages like Hausa, as they can model the sequential dependencies that govern word formation and sentence structure more effectively than models lacking such memory [13, 14]. This finding aligns with prior work demonstrating the efficacy of LSTMs for other morphologically complex, low-resource languages [28, 29].

Furthermore, the strong correlation between word frequency and prediction accuracy (Table 4) is a critical insight. It validates a fundamental principle of data-driven NLP while highlighting the specific challenge of lexical coverage in low-resource settings. The model excels with common vocabulary but requires strategic augmentation to handle rare terms and proper nouns, which are essential for comprehensive usability.

4.8. Contribution to Low-Resource Language NLP

This work makes several concrete contributions to the field. First, it provides a reproducible blueprint for developing user-facing NLP tools in data-scarce environments. The hybrid data strategy merging authentic text with rule-based synthetic generation proves to be a viable solution to the foundational problem of corpus scarcity, a challenge noted across African language NLP [20, 19].

Second, by publicly releasing the curated Hausa dataset, this research actively addresses the resource gap that hinders progress. It contributes to the growing ecosystem of datasets for African languages, such as those for sentiment analysis [21, 22], by providing a resource focused on a generative task.

Most importantly, this study moves beyond purely computational metrics to demonstrate real-world impact through user testing. The significant improvements in typing efficiency and high user acceptance rate translate the technical success into a tangible social benefit, advancing the goal of digital inclusion articulated in the problem statement.

4.9. Limitations and Challenges

While the results are promising, several limitations, informed by the error analysis, must be acknowledged. The performance dip for low-frequency words (Table 4) points to the inherent limitation of a corpus that, despite hybrid generation, cannot fully encapsulate the long-tail of a living language. This is a common challenge for low-resource models [24].

Additionally, the slight decline in accuracy for very long sentences suggests a constraint of the sequential LSTM architecture in modelling extremely long-range dependencies, a weakness that transformer-based models address through self-attention [15]. Our use of an LLM fallback partially mitigates this but does not eliminate the core architectural limitation.

Finally, the scope of the user study, while informative, was limited in scale and demographic diversity. A broader deployment would be necessary to fully understand usability across different age groups, literacy levels, and Hausa dialects (e.g., Kano vs. Sokoto variations).

5. Discussion

This study successfully developed and validated a hybrid neural autocomplete system for the Hausa language, directly addressing the research question of how to build an effective predictive tool despite severe data constraints. The results demonstrate that a strategically designed methodology combining hybrid data generation, a stacked LSTM model, and LLM fallback can yield a system with both high computational accuracy and significant practical utility. This section interprets the key findings, examines their implications for NLP in low-resource contexts, acknowledges limitations, and outlines pathways for future research.

6. Conclusion and Future Research Directions

6.1. Conclusion

This research successfully addressed the critical gap in language technology for Hausa speakers by developing, implementing, and rigorously evaluating a dedicated neural text autocomplete system. Confronting the fundamental challenge of data scarcity inherent to low-resource languages, we proposed and validated a novel methodological framework. This framework integrates a hybrid corpus strategically combining authentic and synthetically generated Hausa text with a stacked Long Short-Term Memory (LSTM) neural network architecture, augmented by a large language model (LLM) for fallback robustness.

The developed system achieved a Top-1 accuracy of 92.4% and a Top-5 accuracy of 98.6%, significantly outperforming traditional statistical and simpler neural baselines. Beyond computational metrics, a user study with 30 participants confirmed its practical utility, demonstrating a 34.3% increase in typing speed and a 91% acceptance rate for primary suggestions. These results empirically validate that tailored machine learning approaches can effectively deliver high-performance digital tools for underserved linguistic communities.

This work makes four primary contributions to the field of natural language processing and digital inclusivity: (1) a reproducible methodology for building NLP applications in low-resource settings; (2) a high-performing, context-aware predictive model for Hausa; (3) a novel, publicly released hybrid Hausa dataset to spur further research; and (4) empirical evidence from user studies quantifying real-world benefits in communication efficiency.

While limitations concerning rare words and long-context modeling persist, this study provides a foundational model and a clear pathway for enhancement through dataset expansion, architectural innovations, and community-driven deployment. By bridging a significant technological divide, this research not only empowers millions of Hausa speakers but also offers a scalable blueprint for supporting other low-resource languages globally, advancing the broader imperative of equitable access to digital communication technologies.

6.2. Future Research Directions

Based on these findings and limitations, we propose the following directions for future work:

- **Advanced Data Augmentation:** To improve rare word prediction, future systems could employ more sophisticated augmentation techniques, such as back-translation using Hausa-English models or contextual synonym replacement within the synthetic generator.
- **Lightweight Transformer Architectures:** Exploring efficient transformer variants (e.g., distilled models or linear transformers) could better handle long-context predictions while remaining feasible for deployment in resource-constrained environments.
- **Dialectal and Code-Switching Integration:** Actively collecting and incorporating text from various Hausa dialects and code-switched (Hausa-English/Arabic) social media data would enhance the model's robustness and real-world applicability.
- **End-to-End System Deployment:** The logical next step is integration into an open-source mobile keyboard application (e.g., as a plugin for OpenBoard or FlorisBoard), enabling large-scale field testing and continuous, privacy-preserving learning from real user interactions.
- **Generalization to Other Languages:** The methodology should be validated on other low-resource African languages to assess its generalizability and refine it into a standard framework for similar tool development.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

- [1] B. Shneiderman and C. Plaisant, *Designing the user interface: strategies for effective human-computer interaction*, Pearson Education India, 2010.
- [2] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.
- [3] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury, "The state and fate of linguistic diversity and inclusion in the NLP world," *arXiv preprint arXiv:2004.09095*, 2020.
- [4] D. M. Eberhard, G. F. Simons, and C. D. Fennig, *Ethnologue: Languages of the world*, SIL International, 2021.
- [5] E. M. Bender and B. Friedman, "Data statements for natural language processing: Toward mitigating system bias and enabling better science," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 587-604, 2018.
- [6] P. Newman, *The Hausa language: An encyclopedic reference grammar*, Yale University Press, 2014.
- [7] B. Caron, "Hausa grammatical sketch," in *The Oxford Handbook of African Languages*, 2019.
- [8] M. Silfverberg, I. S. MacKenzie, and P. Korhonen, "Predicting text entry speed on mobile phones," in *Proc. SIGCHI Conf. on Human Factors in Computing Systems*, pp. 9-16, 2000.
- [9] I. S. MacKenzie and K. Tanaka-Ishii, *Text entry systems: Mobility, accessibility, universality*, Morgan Kaufmann, 2007.
- [10] D. Jurafsky and J. H. Martin, *Speech and language processing (3rd ed. draft)*, 2021. [Online]. Available: <https://web.stanford.edu/~jurafsky/slp3/>.
- [11] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, pp. 3111-3119, 2013.
- [12] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532-1543, 2014.
- [13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997.

- [14] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems*, pp. 3104-3112, 2014.
- [15] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, pp. 5998-6008, 2017.
- [16] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 4171-4186, 2019.
- [17] T. Brown et al., "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877-1901, 2020.
- [18] M. Chen et al., "Smart Compose: Context-aware sentence suggestion in Gmail," in *Proc. 2019 ACM SIGKDD*, 2019.
- [19] A. Conneau et al., "Unsupervised cross-lingual representation learning at scale," in *Proc. 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440-8451, 2020.
- [20] I. Inuwa-Dutse, "NaijaNLP: A survey of Nigerian low-resource languages," *arXiv preprint arXiv:2502.19784*, 2025.
- [21] A. Z. Ibrahim et al., "HausaMovieReview: A benchmark dataset for sentiment analysis in a low-resource African language," *arXiv preprint arXiv:2509.16256*, 2025.
- [22] A. A. Sani, S. H. Muhammad, and D. Jarvis, "Investigating the impact of language-adaptive fine-tuning on sentiment analysis in Hausa using AfriBERTa," *arXiv preprint arXiv:2501.11023*, 2025.
- [23] R. Y. Zakari, Z. K. Lawal, and I. Abdulmumin, "A systematic literature review of Hausa natural language processing," *Int. J. of Computer and Information Technology*, vol. 10, no. 4, pp. 173-186, 2021.
- [24] Z. K. Karami, Z. Lawal, and I. Abdulmumin, "The complexity of Hausa language NLP: Dialects, scripts, and resource constraints," *African Journal of Information Systems*, vol. 13, no. 2, 2021.
- [25] U. Ogbaji and C. N. Ogbuji, "Challenges of orthographic variation in African predominantly oral languages for NLP," *Journal of African Language Technology*, vol. 2, no. 1, pp. 1-20, 2024.
- [26] A. M. Wali and S. Nisioi, "Automatic correction of writing anomalies in Hausa texts," *arXiv preprint arXiv:2506.03820*, 2025.
- [27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [28] C. Baziotis et al., "NTUA-SLP at IEST 2018: Ensemble of neural transfer methods for implicit emotion classification," in *Proc. 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 57-64, 2018.
- [29] R. Sharma, N. Goel, N. Aggarwal, P. Kaur, and C. Prakash, "Next word prediction in Hindi using deep learning techniques," in *Proc. Int. Conf. on Data Science and Engineering (ICDSE)*, pp. 55-60, 2019.