HausaNLP: Current Status, Challenges and Future Directions for Hausa Natural Language Processing

Shamsuddeen Hassan Muhammad[1,5,6], **Ibrahim Said Ahmad**[2,5,6], **Idris Abdulmumin**[3,5], **Falalu Ibrahim Lawan**[4,5], **Sukairaj Hafiz Imam**[5,6], **Yusuf Aliyu**[7], **Sani Abdullahi Sani**[8], **Ali Usman Umar**[9], **Tajuddeen Gwadabe**[5], **Kenneth Church**[2], **Vukosi Marivate**[3]

[1]Imperial College London, [2]Northeastern University, [3]Data Science for Social Impact, University of Pretoria, [4]Kaduna State University, [5]HausaNLP, [6]Bayero University, Kano, [7]Universiti Teknologi PETRONAS, [8]University of the Witwatersrand, Johannesburg, [9]Federal University of Lafia

correspondence: s.muhammad@imperial.ac.uk

Abstract

Hausa Natural Language Processing (NLP) has gained increasing attention in recent years, yet remains understudied as a low-resource language despite having over 120 million first-language (L1) and 80 million second-language (L2) speakers worldwide. While significant advances have been made in high-resource languages, Hausa NLP faces persistent challenges including limited open-source datasets and inadequate model representation. This paper presents an overview of the current state of Hausa NLP, systematically examining existing resources, research contributions, and gaps across fundamental NLP tasks: text classification, machine translation, named entity recognition, speech recognition, and question answering. We introduce HAUSANLP[1], a curated catalog that aggregates datasets, tools, and research works to enhance accessibility and drive further development. Furthermore, we discuss challenges in integrating Hausa into large language models (LLMs), addressing issues of suboptimal tokenization, and dialectal variation. Finally, we propose strategic research directions emphasizing dataset expansion, improved language modeling approaches, and strengthened community collaboration to advance Hausa NLP. Our work provides both a foundation for accelerating Hausa NLP progress and valuable insights for broader multilingual NLP research.

1 Introduction

The limits of my language mean the limits of my world. – (Wittgenstein, 1994)

Natural Language Processing (NLP) has made significant progress and revolutionized the way language technology is used in our daily lives. From voice assistants and chatbots to machine translations, text classification, information extraction, and question-answering, NLP enables us to interact with machines in a more natural way (Cambria and White, 2014). One of the recent advances in NLP is emergence of large language models (LLMs) such as ChatGPT, which demonstrated impressive performance in various NLP tasks, such as dialogue generation and arithmetic reasoning (Qin et al., 2023). However, much of this progress has been concentrated on a limited set of high-resource languages (e.g., English and Chinese), where large-scale pre-training corpora are readily available (van Esch et al., 2022). As a result, many languages remain underrepresented in NLP research, including Hausa.

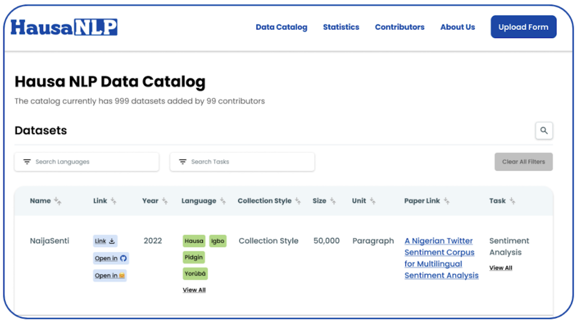

Figure 1: **HausaNLP Catalogue:** A repository of datasets, tools, and research papers on Hausa NLP, developed to improve access to and discovery of Hausa language resources

Hausa is a major Chadic language with rich linguistic and cultural significance within the Afroasiatic family. Originally written in Arabic script (Ajami) during the pre-colonial era, the language has been romanized and now uses the Latin script as its primary writing system. Yet, Arabic influence remains evident in Hausa through loanwords from Arabic (El-Shazly, 1987; Newman, 2022). Most Hausa speakers are found in northern Nigeria and southern Niger. However, its influence has expanded through trade and migration, reaching countries such as Cameroon, Ghana, Benin, Togo, Chad, and Sudan (Inuwa-Dutse, 2023). Hausa has a global presence and is broadcast by several interna-

[1]https://catalog.hausanlp.org

Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025), pages 176–191

tional media outlets such as BBC, Deutsche Welle, Voice of America, Voice of Russia, China Radio International, and Radio France Internationale in Hausa —*the most predominant language broadcast internationally in West Africa*.

Despite its importance, diversity, and cultural heritage, Hausa has received relatively little attention in NLP research (Zakari et al., 2021; Muhammad et al., 2025c; Parida et al., 2023). This slows progress in language technology research and development in Hausa and further widens the gap. Recent work on HausaNLP is mostly community-driven efforts such as machine translation (Adelani et al., 2022a; Abdulmumin et al., 2022b), sentiment analysis (Muhammad et al., 2022, 2023), emotion detection (Muhammad et al., 2025c), hate speech detection (Muhammad et al., 2025a), and named entity recognition (Adelani et al., 2022c). However, numerous NLP tasks for Hausa remain understudied, primarily due to the lack of available corpora.

Open-source corpora are key drivers of advancements in NLP. However, Hausa, a well-documented language, lacks open-source corpora that can be used for many NLP tasks. Further, the few available Hausa corpora are dispersed and difficult to access. Therefore, creating and aggregating open-source corpora for Hausa is crucial for the progress of HausaNLP. To address these challenges, this paper makes the following contributions:

- **HausaNLP Catalogue**: We introduce HausaNLP Catalogue, a centralized repository of datasets, tools, and research papers designed to improve accessibility and accelerate progress in Hausa NLP research.

- **Comprehensive Review**: We present a review of Hausa NLP research, analyzing current progress and identifying key challenges in the field.

- **Future Directions**: We explore promising research opportunities and outline recommendations to advance Hausa NLP technologies.

We release the HausaNLP Catalogue as an open, community-driven platform to centralize and accelerate Hausa NLP research. The catalogue serves as a living resource for discovering and sharing datasets, tools, and papers, with ongoing contributions from researchers and practitioners worldwide.

2 Hausa Language

Hausa is the language of the Hausa people (*Hausawa*), primarily spoken in West Africa's sub-Saharan region, with the largest populations in northern Nigeria and southern Niger. Significant Hausa-speaking communities exist across Northern Ghana, Togo, Cameroon, and parts of Sudan, Chad, Mali, Ivory Coast, Libya, Saudi Arabia, and the Central African Republic (Bello, 2015). With approximately 120 million first-language (L1) and 80 million second-language (L2) speakers, Hausa ranks among Africa's most widely spoken languages, second only to Swahili in total speaker count (Hegazy et al.).

While some argue that Hausa may surpass Swahili in total speakers (Newman, 2022), Swahili maintains broader institutional recognition as an official language in four East African nations: Tanzania, Kenya, Uganda, and Rwanda. In contrast, Hausa had limited official recognition until recently, when Niger declared it an official language (El-Shazly, 1987).

Linguistically, Hausa belongs to the Chadic branch of the Afroasiatic language family and is spoken by over 200 million people either as a first language or as a second language, making it a prominent lingua franca in the region (Yakasai, 2025). Hausa has several dialect variations, which are broadly categorized into two major groups: western and eastern dialects. Furthermore, Hausa has regional variations influenced by contact with non-Hausa languages, leading to phonological, morphological, syntactic, and lexical differences (Bello, 2015).

Phonologically, Hausa is a tonal language with three pitch contrasts that distinguish word meanings and grammatical categories. It has 48 phonemes and 36 standard alphabets (Caron, 2012). Morphologically, Hausa uses root-and-pattern templates and affixation to support complex morphological processes including inflection, derivation, modification, reduplication, clipping, blending, and compounding. It also has numerous loanwords from contact language such as Arabic (Ahmed and B., 1970). Syntactically, Hausa follows a subject-verb-object (SVO) word order and uses diverse typological constructions. The language has developed two writing systems: Ajami (Arabic-based script) and Boko (Latin-based script), both actively used in print, broadcasting, and digital media.

Despite its linguistic richness, Hausa remains a low-resource language in NLP due to limited annotated corpora and tools, hindering the development of language technologies.

3 Current State of Hausa NLP

Several existing works have explored various NLP tasks in Hausa, including text classification, machine translation, named entity recognition, and automatic speech recognition, as shown in Figure 2. This section reviews prior work on Hausa NLP, discusses available datasets, and identifies future research directions.

3.1 Text Classification

Text classification is a method for automatically categorizing texts into distinct, predetermined classes. It is a supervised learning approach, as the classes must be known beforehand to train the model. Text classification can take various forms; however, in the context of Hausa texts, prior studies have primarily focused on sentiment analysis, toxicity detection, or topic classification

Sentiment Analysis Sentiment analysis is a text classification method of categorizing based on the sentiment contained in the text. The method is usually a binary classification, into positive and negative classes, or three classes, into positive, negative, and neutral classes.

Several studies have explored sentiment analysis in Hausa. Abubakar et al. (2021) introduced a sentiment analysis model for Hausa texts, leveraging a corpus of political tweets. Their approach incorporated Hausa lexical features and sentiment intensifiers, achieving an accuracy of 0.71 when employing the SVM classifier. Nevertheless, the dataset size of merely around 200 tweets in the study is grossly inadequate for training supervised learning models.

Muhammad et al. (2022) proposed the first large-scale sentiment dataset for the Hausa language among other Nigerian languages. The paper collected and annotated around 30,000 tweets in the Hausa language. The authors proposed novel methods for tweet collection, filtering, processing, and labeling methods. Additionally, contrary to the other study, they leverage fine-tuning LLMs, attaining a weighted F1-score of 0.81.

Further, Sani et al. (2022) combined machine learning and lexicon-based approaches, achieving 86% accuracy with TF-IDF but struggling with syntactic and semantic nuances. Shehu et al. (2024)

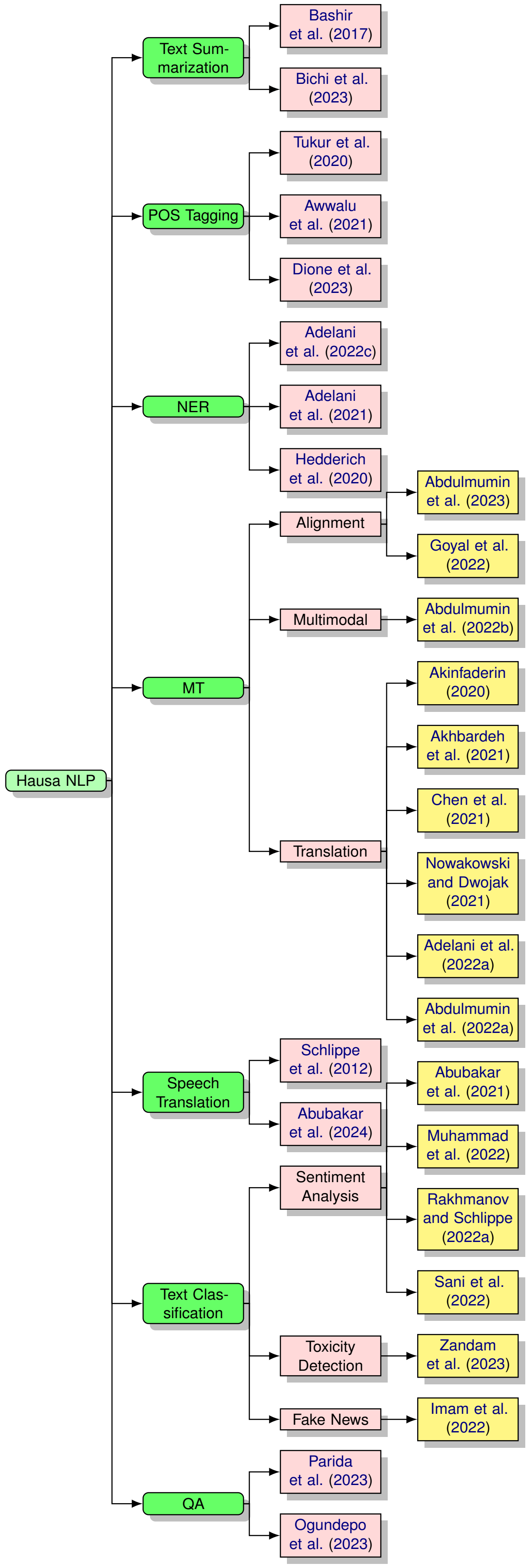

Figure 2: Taxonomy of Hausa NLP Research Progress: Tasks and Associated Publications

integrated CNN, RNN, and HAN with a lexicon dictionary, but the approach yielded a lower accuracy of 68.48%, highlighting the limitations of the bag-of-words model. Mohammed and Prasad (2024) introduced a manually annotated lexicon dataset for social media and product reviews, useful for lexicon-based models but unsuitable for data-driven approaches. To address language-specific challenges, Abdullahi et al. (2024) implemented a normalization process for handling Hausa abbreviations and acronyms, improving the performance of MNB and Logistic Regression. Meanwhile, Ibrahim et al. (2024) proposed a Deep CNN model for aspect and polarity classification in Hausa movie reviews, achieving 92% accuracy but struggling with multi-aspect classification. These studies highlight progress in Hausa sentiment analysis while emphasizing the need for better feature representation, richer datasets, and advanced techniques to handle linguistic complexities.

Future research in Hausa sentiment analysis should focus on high-quality annotated datasets to improve benchmarking (Liu et al., 2024), and domain adaptation to enhance model generalization across different contexts (Hays et al., 2023; Singhal et al., 2023), Cross-lingual sentiment classification offers potential for transferring knowledge from high-resource languages while addressing cultural nuances (Chan et al., 2023; Rakhmanov and Schlippe, 2022b; Yusuf et al., 2024). Further, aspect-based sentiment analysis (ABSA) is crucial for entity-level sentiment detection (Ibrahim et al., 2024; Obiedat et al., 2021), while multimodal approaches integrating text, audio, and visuals remain underexplored (Zhu et al., 2023; Gandhi et al., 2023; Parida et al., 2023). Sentiment analysis using code-mixed remains underexplored in HausaNLP (Shakith and Arockiam, 2024; Yusuf et al., 2023). Finally, explainable sentiment analysis should be explored to improve model transparency (Diwali et al., 2023). Advancing these areas will significantly strengthen Hausa NLP research and applications.

Emotion analysis in text Unlike sentiment analysis, which aims to interpret text and assign polarities (positive, negative, or neutral), emotion analysis focuses on extracting and analyzing fine-grained emotions, known as affects (e.g., happiness, sadness, fear, anger, surprise, and disgust). Muhammad et al. (2025b) is the first work on emotion detection in Hausa. The authors developed a text-based emotion dataset in 29 languages, including Hausa. The dataset is annotated into six emotion classes (anger, fear, joy, sadness, surprise, and disgust) and further categorized into intensity levels: 0 (indicating no emotion), 1 (low emotion), 2 (medium emotion), and 3 (high emotion). This dataset was used in the SemEval shared task (Muhammad et al., 2025b).

Toxicity detection Toxicity detection is a text classification task of detecting toxicity in text. The toxicity could be in the form of hate speech, harassment, and threats. The only work on toxicity detection in Hausa texts is by (Zandam et al., 2023). In the work, the authors developed an online threat detection dataset using both Facebook and Twitter posts. The developed dataset is quite limited with around 801 instances. The Hausa threat detection models are based on machine learning algorithms, achieving the best performance of 0.85 with a random forest algorithm.

Fake news detection The advancement of the internet and social media has accelerated news dissemination, offering both benefits and drawbacks. While crucial information reaches the public swiftly, the downside includes the widespread circulation of fake news. It is increasingly become difficult to distinguish actual news and fake news in the cyberspace. As a result, fake news detection has become an important area of research.

The work of Imam et al. (2022) focused on the creation of fake news detection corpus for Hausa news articles. They developed a corpus of 2600 news articles comprising of real and fake news selected from key topics like: Business, health, entertainment, sports, politics and religion.

Topic Classification News topic classification is a text classification task in NLP that involves categorizing news articles into different categories like sports, business, entertainment, and politics. For Hausa news articles, Adelani et al. (2023) focused on topic classification for African langauges' news articles including Hausa articles. They used both classical machine learning algorithms, and pre-trained LLMs. The best performing model is AfroXLMR-large attaining a weighted F1-score of 0.92.

3.2 Machine Translation

3.2.1 Text Translation

Adelani et al. (2022a) leveraged pre-trained models for African news translation, focusing on 16 under-represented African languages including the Hausa language. For the Hausa language, The Hausa Khamenei [2] corpus contained 5,898 sentences, was used. The study demonstrated the effectiveness of fine-tuning pre-trained models on a few thousand high-quality bitext for adding new languages like Hausa to the models.

Nowakowski and Dwojak (2021) and Chen et al. (2021) participated in the WMT 2021 News Translation Task (Akhbardeh et al., 2021). This involves building a machine translation system for English and Hausa language pairs. The Nowakowski and Dwojak (2021) focused on thorough data cleaning, transfer learning, iterative training, and back-translation. The work experimented with NMT and PB-SMT, using the base Transformer architecture for the NMT models. On the other hand, (Chen et al., 2021) used an iterative back-translation approach on top of pre-trained English-German models and investigated vocabulary embedding mapping.

Akinfaderin (2020) explored English-Hausa machine translation by training LSTM and transformer-based model using the JW300 (Agić and Vulić, 2019) corpus. Abdulmumin et al. (2022a) participated in WMT 2022 Large-Scale Machine Translation Evaluation for the African Languages Shared Task (Adelani et al., 2022b). The work made an attempt to improve Hausa-English (along with other language pairs) machine translation using data filtering techniques. The idea relies on filtering out the noisy or invalid parts of a large corpus, keeping only a high-quality subset thereof. The results show that the performance of the models improved with increased data filtering, indicating the removal of noisy sentences enhanced translation quality.

3.2.2 Multi-Modal Machine Translation

Multimodal machine translation (MMT) focuses on translating languages using multiple modalities of information, not just text. This typically involves combining text with other data sources, such as images, speech, and video. MMT aims to enhance translation quality by incorporating information from other modalities. The goal is to leverage these additional modalities to improve the overall translation process.

Abdulmumin et al. (2022b) presents the ***Hausa Visual Genome (HaVG)***, a multi-modal dataset that contains the description of an image or a section within the image in Hausa and its equivalent in English. HaVG was formed by translating the English description of the images in the Hindi Visual Genome (HVG) into Hausa automatically. Afterward, the synthetic Hausa data was carefully post-edited considering the respective images. The dataset comprises 32,923 images and their descriptions.

3.2.3 Sentence Alignment

Automatic sentence alignment is the process of identifying which sentences in a source text correspond to which sentences in a target text. This task is crucial for creating parallel corpora, where each sentence in one language is aligned with its equivalent translation in another language. Various approaches, including length-based, lexicon-based, and translation-based methods, are employed for sentence alignment. Evaluating alignment quality involves assessing accuracy and effectiveness, considering factors like language pairs and genre.

Abdulmumin et al. (2023) addresses the challenge of limited qualitative datasets for English-Hausa machine translation by automatic sentence alignment. The work presented a qualitative parallel sentence aligner that leverages the closed-access Cohere multilingual embedding [3]. For evaluation, the work used the MAFAND-MT (Adelani et al., 2022a), FLORES (Goyal et al., 2022), a new corpus of 1000 Hausa and English news articles each. The proposed method showed promising results.

3.3 POS

Part-of-speech tagging (POS) is one of the first steps in NLP that involves the tagging (or labeling) of each word in a sentence with the correct part of speech to indicate their grammatical behaviours for computational tasks (Martinez, 2012). POS tagging is very crucial in many NLP tasks like sentiment analysis and information extraction.

While considerable amount of work has been done on POS tagging, only a couple of studies are on Hausa POS tagging. Tukur et al. (2020) proposed a technique for POS tagging of Hausa

[2] https://www.statmt.org/wmt21/translation-task.html

[3] https://docs.cohere.com/docs/multilingual-language-models

sentences using the Hidden Markov Model. They evaluated the model using a manually collected and annotated Hausa corpus sourced from from radio stations. While the study is worthwhile, both the dataset and model are not publicly available.

Awwalu et al. (2021) presents a study on Corpus Based Transformation-Based Learning for Hausa language POS tagging. The research involves corpus development for Hausa language POS tagset. Various models and techniques such as Transformation-Based Learning (TBL), Hidden Markov Model (HMM), and N-Gram models are employed for POS tagging. The main findings indicate that the TBL tagger outperforms HMM and N-Gram taggers in terms of accuracy levels, showcasing the effectiveness of hybrid generative and discriminative taggers.

Dione et al. (2023) created MasakhaPOS, a large POS dataset for 20 diverse African languages. They address the challenges of using universal dependencies (UD) guidelines for these languages, and compare different POS taggers based Conditional Random Field (CRF) and several multilingual Pretrained Language Models (PLMs). For the Hausa part of the project, the data was sourced from *Kano Focus* and *Freedom Radio* to a total of 1504 sentences (train: 753, test:150, and dev: 601).

3.4 Text Summarization

Text summarization is the process of automatically generating a concise and coherent summary of a longer text while retaining its key information and main points (El-Kassas et al., 2021).

Text summarization plays a crucial role in various applications such as information retrieval, document summarization, news aggregation, and content recommendation systems, helping users quickly grasp the main points of lengthy documents or articles.

(Bashir et al., 2017) perhaps conducted one the the earliest works on text summarization for Hausa langauge. The work focused on text summarization based on feature extraction using Naive Bayes model. However, the validity of the work is limited by the small data size of 10 documents from news articles, with each document containing over 600 words. The work of (Bichi et al., 2023) focus on graph-based extractive text summarization method for Hausa text. The study focus on graph-based extractive single-document summarization method for Hausa text by modifying the PageRank algorithm using the normalized common bigrams count between adjacent sentences as the initial vertex score. They evaluated the proposed approach using a manually annotated dataset that comprises of 113 Hausa news articles on various genres. Each news article had two manually generated gold standard summaries, with the length of summaries being 20% of the original article length.

3.5 Question and Answering

Question and Answering (QA) is a branch of natural language processing (NLP) that deals with building systems that can automatically answer questions posed by humans in natural language. QA systems can be useful for various applications, such as virtual assistants, customer support, search engines, and education (Rogers et al., 2023).

Parida et al. (2023) developed a Hausa Visual Question Answering (VQA) dataset called ***HaVQA***. The dataset is a multi-modal dataset for visual question-answering (VQA) tasks in the Hausa language. The dataset was created by manually translating 6,022 English question-answer pairs, which are associated with 1,555 unique images from the Visual Genome dataset. The paper employed state-of-the-art language and vision models for Visual Question Answering and achieved the best performance with the Data-Efficient Image Transformers model proposed by Facebook with a WuPalmer score of 30.85.

(Ogundepo et al., 2023) developed ***AfriQA***, a dataset for cross-lingual open-retrieval question answering for 10 African languages, including the Hausa language. The dataset was developed from Wikipedia articles and manually elicited questions. For Hausa language, the final corpus consist of 1171 instances split into 435 training, 436 development and 300 test sets. The findings of the experiments proves how challenging multilingual retrieval is even for state-of-the-art QA models.

3.6 Named Entity Recognition

Named entity recognition (NER) is a technique of NLP that identifies and classifies named entities in a text, such as person names, organizations, locations, and dates. NER can be useful for various tasks, such as information extraction, search engines, chatbots, and machine translation. There are different methods and tools for NER, such as dictionary-based, rule-based, machine learning-based, and hybrid systems (Li et al., 2022).

Adelani et al. (2021) and Adelani et al. (2022c) created the largest NER corpus for African languages titled ***MasakhaNER 1.0*** and ***MasakhaNER 2.0***. MasakhaNER 1.0 covers 10 African languages, while MasakhaNER 2.0 expanded the corpus to include 10 South African languages, making a total of 20 languages. MasakhaNER 1.0 consists of 2,720 sources from VOA news while MasakhaNER 2.0 consists of 8,165 sourced from Kano Focus and Freedom Radio news channels. Both studies explored various experiments using pretrained language models and other techniques like transfer learning and zero-shot learning.

The work of Hedderich et al. (2020) investigates transfer learning and distant supervision with multilingual transformer models on NER and topic classification in Hausa, isiXhosa and Yoruba languages. The study show that transfer learning from a high-resource language and distant supervision are effective techniques for improving performance in low-resource settings for African languages.

3.7 Automatic Speech Recognition (ASR)

Automatic speech recognition (ASR) is a technology that allows computers to convert spoken language into text. ASR can be used for various purposes, such as voice control, transcription, translation, and accessibility (Yu and Deng, 2016).

Schlippe et al. (2012) focused on developing a Hausa Large Vocabulary Continuous Speech Recognition (LVCSR) system by collecting a corpus of Hausa speech data from native speakers in Cameroon and text data from prominent Hausa websites. The data collected for the study included approximately 8 hours and 44 minutes of speech data from 102 native speakers of Hausa in Cameroon. Additionally, the text corpus consists of roughly 8 million words. The study found that modeling tones and vowel lengths significantly improved recognition performance, leading to a reduction in word error rates.

(Abubakar et al., 2024) focuses on developing a diacritic-aware automatic speech recognition model for the Hausa language. The model uses a large corpus of speech data from the Mozilla Common Voice dataset, which includes a variety of diacritical words and sentences. The Whisper-large model outperforms existing models, achieving a word error rate of 4.23% and a diacritic coverage of 92%. It also has a precision of 98.87%, with a 2.1% diacritic error rate, demonstrating its effectiveness in accurately transcribing Hausa speech. However, Due to the absence of prior ASR systems specifically focused on diacritization in the Hausa language, the authors were unable to make direct comparisons with their results. This lack of benchmarks may limit the ability to fully assess the effectiveness of their proposed model against existing technologies

Future efforts should prioritize developing real-time ASR systems for continuous Hausa speech recognition, enhancing usability across everyday communication and diverse industries. Optimizing computational resources and designing efficient algorithms will enable high-performance ASR systems with reduced power requirements. Further, exploring ASR techniques less reliant on diacritics can broaden usability for varied contexts and users. Finally, integrating ASR with NLP and machine translation can pave the way for comprehensive tools to better serve Hausa-speaking communities.

4 Hausa Representation in Large Language Models (LLMs)

Large language models (LLMs) have made significant strides in supporting multilingual tasks, including those involving low-resource languages like Hausa. Multilingual models such as AfrIBERTa (Ogueji et al., 2021) mBERT (Devlin et al., 2019), InkubaLM (Tonja et al., 2024) XLM-R (Conneau et al., 2020), and BLOOM (Workshop et al., 2023) have incorporated Hausa into their training data, albeit to varying degrees. These models leverage cross-lingual transfer learning to improve performance on languages with limited resources. However, the extent of Hausa representation in these models is often constrained by the scarcity of high-quality, diverse datasets.

The availability and quality of training data are critical factors influencing the performance of large language models (LLMs) on Hausa language tasks. Like many low-resource languages, Hausa faces challenges such as data scarcity, representational bias, and inadequate dataset construction. Existing datasets are often limited in scale and diversity, particularly in capturing dialectal variations and informal text (e.g., social media content). Sani et al. (2025b) highlight these challenges, emphasizing the impact of dialectal variation and tokenization on Hausa sentiment analysis. Their findings underscore the need for more diverse and high-quality datasets to enhance model performance. Without sufficient data, LLMs struggle to achieve robust performance in handling Hausa text, as highlighted by Zhao et al. (2024) and Acikgoz et al. (2024).

In addition to data scarcity, Hausa's linguistic features pose significant challenges for tokenization and language modeling. The language's rich morphology, tonal variations, and complex noun pluralization systems complicate the process of accurately representing it in LLMs. Diacritics and tonal markers, which are critical for meaning, often lead to suboptimal tokenization, resulting in poor representations of the language (Abubakar et al., 2024; Jaggar, 2006). Furthermore, the dialectal diversity within Hausa adds another layer of complexity. Models trained on formal Hausa text frequently struggle to process informal or dialectal variations, as noted by Sani et al. (2025b). This limits their applicability in real-world scenarios where such variations are common.

Another critical issue is bias and representation in existing LLMs. Studies comparing LLM outputs with native speaker responses have revealed discrepancies in how cultural nuances and emotional tones are captured (Ahmad et al., 2024). These biases can lead to outputs that are misaligned with the cultural and linguistic expectations of Hausa speakers, further reducing the utility of LLMs for this language. Addressing these challenges requires innovative approaches, including improved tokenization strategies, dialectal adaptation techniques, and data augmentation methods. By tackling these issues, researchers can develop more robust and inclusive models that better serve Hausa speakers and other low-resource language communities

A promising direction is the development of specialized, lightweight models tailored specifically to Hausa. These custom models could provide more accurate and efficient solutions for Hausa-specific applications (Yang et al., 2024). Additionally, federated prompt tuning offers a pathway to enhance data efficiency and facilitate mutual improvements across languages, benefiting low-resource languages like Hausa (Zhao et al., 2024). Synthetic data generation also presents a valuable opportunity to address data scarcity. By creating high-quality synthetic datasets, researchers can overcome the limitations of limited real-world data and improve the performance of the model (Mahgoub et al., 2024). Together, these approaches, ranging from architectural innovations and specialized models to federated learning and synthetic data, have the potential to significantly advance Hausa representation in LLMs, making them more robust, efficient, and culturally relevant for Hausa speakers.

5 Conclusion

Advancing Hausa NLP requires a multifaceted approach that addresses both technical and community-driven challenges. Below, we outline key areas for future research and development.

Future research should investigate the interplay between tokenization strategies and model initialization to optimize the learning efficiency of Hausa LLMs. Techniques inspired by the BabyLM Challenge (Hu et al., 2024) could be adapted to Hausa, focusing on sample-efficient pretraining and developmentally plausible corpora. Such approaches could mitigate data scarcity while improving model performance, particularly in low-resource settings.

Innovative architectures that support dynamic retokenization based on context could significantly enhance the representation of Hausa's linguistic features. These models would adapt tokenization to better capture dialectal variations and morphological complexity, improving generalization across diverse Hausa texts. This is especially important given the language's rich morphology and tonal variations, which are often underrepresented in current models.

Building on the work of Wolf et al. (2023), future studies could explore encoding prosodic features into embeddings to improve the contextual understanding of Hausa. Although prosody carries information beyond text, its integration could enhance model performance, particularly in low-resource settings. This approach could also facilitate better handling of tonal variations in Hausa, which are critical for accurate language representation.

Creating richer and more diverse datasets for Hausa is essential for advancing NLP applications. Future efforts should focus on curating datasets that capture both formal and informal text, as well as dialectal variations. Techniques such as data augmentation, synthetic data generation, and crowdsourcing could help address data scarcity and improve model robustness. Expanding digital resources through initiatives like web crawling and community contributions (Schlippe et al., 2012; Ibrahim et al., 2022) will also play a crucial role.

Engaging the Hausa-speaking community in dataset creation and model evaluation is vital for ensuring that LLMs reflect the linguistic and cultural nuances of Hausa. Collaborative efforts between researchers, linguists, and native speakers could lead to more representative and inclusive models. Community-driven approaches can also help address biases and improve the cultural and emotional representation of Hausa in NLP systems (Ahmad et al., 2024).

Multilingual and cross-lingual transfer learning offers promising opportunities to leverage resources from related languages to enhance Hausa NLP. For instance, the work of Erasmo Ndomba et al. (2025) demonstrates that language-specific tokenizers outperform multilingual tokenizers in tasks like sentiment and news classification for African languages. Interestingly, their findings reveal that a tokenizer trained on Swahili outperformed one trained on Hausa for Hausa-specific tasks, highlighting strong cross-linguistic connections between these languages. This suggests that shared linguistic structures and features among African languages can be harnessed to improve model performance. Future research should explore these cross-linguistic bonds further, leveraging multilingual capabilities and federated learning techniques to enhance Hausa NLP (Zhao et al., 2024).

Adapting and fine-tuning existing LLMs to better handle the unique linguistic features of Hausa is another critical area for future work (Acikgoz et al., 2024; Abubakar et al., 2024). Additionally, addressing biases and ensuring culturally aware models will be essential for creating systems that accurately represent the emotions and nuances of the Hausa language (Ahmad et al., 2024).

References

Habeeba Ibraheem Abdullahi, Muhammad Aminu Ahmad, and Khalid Haruna. 2024. Twitter sentiment analysis for hausa abbreviations and acronyms. *Science World Journal*, 19(1):101–104.

Idris Abdulmumin, Michael Beukman, Jesujoba Alabi, Chris Chinenye Emezue, Everlyn Chimoto, Tosin Adewumi, Shamsuddeen Muhammad, Mofetoluwa Adeyemi, Oreen Yousuf, Sahib Singh, and Tajuddeen Gwadabe. 2022a. Separating grains from the chaff: Using data filtering to improve multilingual translation for low-resourced African languages. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1001–1014, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Idris Abdulmumin, Satya Ranjan Dash, Musa Abdullahi Dawud, Shantipriya Parida, Shamsuddeen Muhammad, Ibrahim Sa'id Ahmad, Subhadarshi Panda, Ondřej Bojar, Bashir Shehu Galadanci, and Bello Shehu Bello. 2022b. Hausa visual genome: A dataset for multi-modal English to Hausa machine translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6471–6479, Marseille, France. European Language Resources Association.

Idris Abdulmumin, Auwal Abubakar Khalid, Shamsuddeen Hassan Muhammad, Ibrahim Said Ahmad, Lukman Jibril Aliyu, Babangida Sani, Bala Mairiga Abduljalil, and Sani Ahmad Hassan. 2023. Leveraging closed-access multilingual embedding for automatic sentence alignment in low resource languages.

Abdulqahar Mukhtar Abubakar, Deepa Gupta, and Susmitha Vekkot. 2024. Development of a diacritic-aware large vocabulary automatic speech recognition for hausa language. *International Journal of Speech Technology*, 27(3):687–700.

Amina Imam Abubakar, Abubakar Roko, Aminu Muhammad Bui, and Ibrahim Saidu. 2021. An enhanced feature acquisition for sentiment analysis of english and hausa tweets. *International Journal of Advanced Computer Science and Applications*, 12(9).

Emre Can Acikgoz, Mete Erdogan, and Deniz Yuret. 2024. Bridging the bosphorus: Advancing turkish large language models through strategies for low-resource language adaptation and benchmarking. page 242 – 268.

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, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022a. 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 Adelani, Md Mahfuz Ibn Alam, Antonios Anastasopoulos, Akshita Bhagia, Marta R. Costa-jussà, Jesse Dodge, Fahim Faisal, Christian Federmann, Natalia Fedorova, Francisco Guzmán, Sergey Koshelev, Jean Maillard, Vukosi Marivate, Jonathan Mbuya, Alexandre Mourachko, Safiyyah Saleem, Holger Schwenk, and Guillaume Wenzek. 2022b. Findings of the WMT'22 shared task on large-scale machine translation evaluation for African languages. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 773–800, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo Lerato Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Oluwaseun Adeyemi, Gilles Quentin Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu, and Dietrich Klakow. 2022c. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. 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, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure F. P. Dossou, Akintunde Oladipo, Doreen Nixdorf, Chris Chinenye Emezue, sana al azzawi, Blessing Sibanda, Davis David, Lolwethu Ndolela, Jonathan Mukiibi, Tunde Ajayi, Tatiana Moteu, Brian Odhiambo, Abraham Owodunni, Nnaemeka Obiefuna, Muhidin Mohamed, Shamsuddeen Hassan Muhammad, Teshome Mulugeta Ababu, Saheed Abdullahi Salahudeen, Mesay Gemeda Yigezu, Tajuddeen Gwadabe, Idris Abdulmumin, Mahlet Taye, Oluwabusayo Awoyomi, Iyanuoluwa Shode, Tolulope Adelani, Habiba Abdulganiyu, Abdul-Hakeem Omotayo, Adetola Adeeko, Abeeb Afolabi, Anuoluwapo Aremu, Olanrewaju Samuel, Clemencia Siro, Wangari Kimotho, Onyekachi Ogbu, Chinedu Mbonu, Chiamaka Chukwuneke, Samuel Fanijo, Jessica Ojo, Oyinkansola Awosan, Tadesse Kebede, Toadoum Sari Sakayo, Pamela Nyatsine, Freedmore Sidume, Oreen Yousuf, Mardiyyah Oduwole, Tshinu Tshinu, Ussen Kimanuka, Thina Diko, Siyanda Nxakama, Sinodos Nigusse, Abdulmejid Johar, Shafie Mohamed, Fuad Mire Hassan, Moges Ahmed Mehamed, Evrard Ngabire, Jules Jules, Ivan Ssenkungu, and Pontus Stenetorp. 2023. Masakhanews: News topic classification for african languages.

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Ibrahim Said Ahmad, Shiran Dudy, Resmi Ramachandranpillai, and Kenneth Church. 2024. Are generative language models multicultural? a study on hausa culture and emotions using chatgpt. page 98 – 106.

U. Ahmed and Dauda B. 1970. An introduction to classical hausa and major dialects. *Norther Nigeria Publishing Company*.

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages

1–88, Online. Association for Computational Linguistics.

Adewale Akinfaderin. 2020. Hausamt v1. 0: Towards english–hausa neural machine translation. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, pages 144–147.

Saminu Mohammad Aliyu, Gregory Maksha Wajiga, Muhammad Murtala, Shamsuddeen Hassan Muhammad, Idris Abdulmumin, and Ibrahim Said Ahmad. 2022. Herdphobia: A dataset for hate speech against fulani in nigeria. In *Seventh Widening Natural Language Processing Workshop (WiNLP)*.

Jamilu Awwalu, Saleh Elyakub Abdullahi, and Abraham Eseoghene Evwiekpaefe. 2021. A corpus based transformation-based learning for hausa text parts of speech tagging. *International Journal of Computing and Digital Systems*, 10:473–490.

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

A. Bello. 2015. The dialects of hausa. *Ahmadu Bello University Press.*

Abdulkadir Abubakar Bichi, Ruhaidah Samsudin, Rohayanti Hassan, Layla Rasheed Abdallah Hasan, and Abubakar Ado Rogo. 2023. Graph-based extractive text summarization method for hausa text. *Plos one*, 18(5):e0285376.

Erik Cambria and Bebo White. 2014. Jumping nlp curves: A review of natural language processing research. *IEEE Computational intelligence magazine*, 9(2):48–57.

Bernard Caron. 2012. Hausa grammatical sketch.

Jireh Yi-Le Chan, Khean Thye Bea, Steven Mun Hong Leow, Seuk Wai Phoong, and Wai Khuen Cheng. 2023. State of the art: a review of sentiment analysis based on sequential transfer learning. *Artificial Intelligence Review*, 56(1):749–780.

Pinzhen Chen, Jindřich Helcl, Ulrich Germann, Laurie Burchell, Nikolay Bogoychev, Antonio Valerio Miceli Barone, Jonas Waldendorf, Alexandra Birch, and Kenneth Heafield. 2021. The University of Edinburgh's English-German and English-Hausa submissions to the WMT21 news translation task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 104–109, Online. Association for Computational Linguistics.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

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, Catherine Gitau, Derguene Mbaye, Jonathan Mukiibi, Blessing Sibanda, Bonaventure F. P. Dossou, Andiswa Bukula, Rooweither Mabuya, Allahsera Auguste Tapo, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Fatoumata Ouoba Kabore, Amelia Taylor, Godson Kalipe, Tebogo Macucwa, Vukosi Marivate, Tajuddeen Gwadabe, Mboning Tchiaze Elvis, Ikechukwu Onyenwe, Gratien Atindogbe, Tolulope Adelani, Idris Akinade, Olanrewaju Samuel, Marien Nahimana, Théogène Musabeyezu, Emile Niyomutabazi, Ester Chimhenga, Kudzai Gotosa, Patrick Mizha, Apelete Agbolo, Seydou Traore, Chinedu Uchechukwu, Aliyu Yusuf, Muhammad Abdullahi, and Dietrich Klakow. 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, Toronto, Canada. Association for Computational Linguistics.

Arwa Diwali, Kawther Saeedi, Kia Dashtipour, Mandar Gogate, Erik Cambria, and Amir Hussain. 2023. Sentiment analysis meets explainable artificial intelligence: A survey on explainable sentiment analysis. *IEEE Transactions on Affective Computing*.

Wafaa S. El-Kassas, Cherif R. Salama, Ahmed A. Rafea, and Hoda K. Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679.

Mohamed Helal Ahmed Sheref El-Shazly. 1987. *The provenance of Arabic loan-words in Hausa: a phonological and semantic study*. University of London, School of Oriental and African Studies (United Kingdom).

Goodwill Erasmo Ndomba, Medard Edmund Mswahili, and Young-Seob Jeong. 2025. Tokenizers for african languages. *IEEE Access*, 13:1046–1054.

Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. 2023. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, 91:424–444.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Chris Hays, Zachary Schutzman, Manish Raghavan, Erin Walk, and Philipp Zimmer. 2023. Simplistic collection and labeling practices limit the utility of benchmark datasets for twitter bot detection. In *Proceedings of the ACM web conference 2023*, pages 3660–3669.

Michael A. Hedderich, David Adelani, Dawei Zhu, Jesujoba Alabi, Udia Markus, and Dietrich Klakow. 2020. Transfer learning and distant supervision for multilingual transformer models: A study on African languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2580–2591, Online. Association for Computational Linguistics.

Mahmoud Fahmi Hegazy, Mohammad Ali Nofal, and MA Mahmoud Sayed. A lexical semantic error analysis of arabic-speaking hausa language learners.

Michael Y. Hu, Aaron Mueller, Candace Ross, Adina Williams, Tal Linzen, Chengxu Zhuang, Ryan Cotterell, Leshem Choshen, Alex Warstadt, and Ethan Gotlieb Wilcox. 2024. Findings of the second BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 1–21, Miami, FL, USA. Association for Computational Linguistics.

Umar Ibrahim, Abubakar Yakubu Zandam, Fatima Muhammad Adam, and Aminu Musa. 2024. A deep convolutional neural network-based model for aspect and polarity classification in hausa movie reviews. *arXiv preprint arXiv:2405.19575*.

Umar Adam Ibrahim, Moussa Boukar Mahatma, and Muhammed Aliyu Suleiman. 2022. Framework for hausa speech recognition.

Sukairaj Hafiz Imam, Abubakar Ahmad Musa, and Ankur Choudhary. 2022. The first corpus for detecting fake news in hausa language. In *Emerging Technologies for Computing, Communication and Smart Cities*, pages 563–576, Singapore. Springer Nature Singapore.

Isa Inuwa-Dutse. 2023. The first large scale collection of diverse hausa language datasets. In *4th Workshop on African Natural Language Processing*.

P.J. Jaggar. 2006. *Hausa*.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2022. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Jiabei Liu, Keqin Li, Armando Zhu, Bo Hong, Peng Zhao, Shuying Dai, Changsong Wei, Wenqian Huang, and Honghua Su. 2024. Application of deep learning-based natural language processing in multilingual sentiment analysis. *Mediterranean Journal of Basic and Applied Sciences (MJBAS)*, 8(2):243–260.

Abeer Mahgoub, Ghada Khoriba, and Elhassan Anas Elsabry. 2024. Mathematical problem solving in arabic: Assessing large language models. volume 244, page 86 – 95.

Angel R. Martinez. 2012. Part-of-speech tagging. *WIREs Computational Statistics*, 4(1):107–113.

Idi Mohammed and Rajesh Prasad. 2024. Lexicon dataset for the hausa language. *Data in Brief*, 53:110124.

Shamsuddeen Muhammad, Idris Abdulmumin, Abinew Ayele, Nedjma Ousidhoum, David Adelani, Seid Yimam, Ibrahim Ahmad, Meriem Beloucif, Saif Mohammad, Sebastian Ruder, Oumaima Hourrane, Alipio Jorge, Pavel Brazdil, Felermino Ali, Davis David, Salomey Osei, Bello Shehu-Bello, Falalu Lawan, Tajuddeen Gwadabe, Samuel Rutunda, Tadesse Belay, Wendimu Messelle, Hailu Balcha, Sisay Chala, Hagos Gebremichael, Bernard Opoku, and Stephen Arthur. 2023. AfriSenti: A Twitter sentiment analysis benchmark for African languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13968–13981, Singapore. Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, David Ifeoluwa Adelani, Ibrahim Said Ahmad, Saminu Mohammad Aliyu, Nelson Odhiambo Onyango, Lilian D. A. Wanzare, Samuel Rutunda, Lukman Jibril Aliyu, Esubalew Alemneh, Oumaima Hourrane, Hagos Tesfahun Gebremichael, Elyas Abdi Ismail, Meriem Beloucif, Ebrahim Chekol Jibril, Andiswa Bukula, Rooweither Mabuya, Salomey Osei, Abigail Oppong, Tadesse Destaw Belay, Tadesse Kebede Guge, Tesfa Tegegne Asfaw, Chiamaka Ijeoma Chukwuneke, Paul Rottger, Seid Muhie Yimam, and Nedjma Djouhra Ousidhoum. 2025a. Afrihate: A multilingual collection of hate speech and abusive language datasets for african languages. *ArXiv*, abs/2501.08284.

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, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, Guanqun Cao, Rodrigo Tufino Cardenas, Rendi Chevi, Chiamaka Ijeoma Chukwuneke, Alexandra Ciobotaru, Daryna Dementieva, Murja Sani Gadanya, Robert Geislinger, Bela Gipp, Oumaima Hourrane, Oana Ignat, Falalu Ibrahim Lawan, Rooweither Mabuya, Rahmad Mahendra, Vukosi Marivate, Andrew Piper, Alexander Panchenko, Charles Henrique Porto Ferreira, Vitaly Protasov, Samuel Rutunda, Manish Shrivastava, Aura Cristina Udrea, Lilian Diana Awuor Wanzare, Sophie Wu, Florian Valentin Wunderlich, Hanif Muhammad Zhafran, Tianhui Zhang, Yi Zhou, and Saif M. Mohammad. 2025b. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, et al. 2025c. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages. *arXiv preprint arXiv:2502.11926*.

Paul Newman. 2022. *Loanwords*, page 205–211. Cambridge University Press.

Artur Nowakowski and Tomasz Dwojak. 2021. Adam Mickiewicz University's English-Hausa submissions to the WMT 2021 news translation task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 167–171, Online. Association for Computational Linguistics.

Ruba Obiedat, Duha Al-Darras, Esra Alzaghoul, and Osama Harfoushi. 2021. Arabic aspect-based sentiment analysis: A systematic literature review. *IEEE Access*, 9:152628–152645.

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.

Odunayo Ogundepo, Tajuddeen Gwadabe, Clara Rivera, Jonathan Clark, Sebastian Ruder, David Adelani, Bonaventure Dossou, Abdou Diop, Claytone Sikasote, Gilles Hacheme, Happy Buzaaba, Ignatius Ezeani, Rooweither Mabuya, Salomey Osei, Chris Emezue, Albert Kahira, Shamsuddeen Muhammad, Akintunde Oladipo, Abraham Owodunni, Atnafu Tonja, Iyanuoluwa Shode, Akari Asai, Anuoluwapo Aremu, Ayodele Awokoya, Bernard Opoku, Chiamaka Chukwuneke, Christine Mwase, Clemencia Siro, Stephen Arthur, Tunde Ajayi, Verrah Otiende, Andre Rubungo, Boyd Sinkala, Daniel Ajisafe, Emeka Onwuegbuzia, Falalu Lawan, Ibrahim Ahmad, Jesujoba Alabi, Chinedu Mbonu, Mofetoluwa Adeyemi, Mofya Phiri, Orevaoghene Ahia, Ruqayya Iro, and Sonia Adhiambo. 2023. Cross-lingual open-retrieval question answering for African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14957–14972, Singapore. 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.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Ochilbek Rakhmanov and Tim Schlippe. 2022a. 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.

Ochilbek Rakhmanov and Tim Schlippe. 2022b. 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.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *ACM Comput. Surv.*, 55(10).

Babangida Sani, Aakansha Soy, Sukairaj Hafiz Imam, Ahmad Mustapha, Lukman Jibril Aliyu, Idris Abdulmumin, Ibrahim Said Ahmad, and Shamsuddeen Hassan Muhammad. 2025a. Who wrote this? identifying machine vs human-generated text in hausa. *arXiv preprint arXiv:2503.13101*.

Muhammad Sani, Abubakar Ahmad, and Hadiza S Abdulazeez. 2022. Sentiment analysis of hausa language tweet using machine learning approach. *Journal of Research in Applied Mathematics*, 8(9):07–16.

Sani Abdullahi Sani, Shamsuddeen Hassan Muhammad, and Devon Jarvis. 2025b. 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.

Tim Schlippe, Edy Guevara Komgang Djomgang, Ngoc Thang Vu, Sebastian Ochs, and Tanja Schultz. 2012. Hausa large vocabulary continuous speech recognition. In *Spoken Language Technologies for Under-Resourced Languages*.

Ayesha Shakith and L Arockiam. 2024. Enhancing classification accuracy on code-mixed and imbalanced data using an adaptive deep autoencoder and xgboost. *The Scientific Temper*, 15(03):2598–2608.

Harisu Abdullahi Shehu, Kaloma Usman Majikumna, Aminu Bashir Suleiman, Stephen Luka, Md Haidar Sharif, Rabie A Ramadan, and Huseyin Kusetogullari. 2024. Unveiling sentiments: A deep dive into sentiment analysis for low-resource languages–a case study on hausa texts. *IEEE Access*.

Peeyush Singhal, Rahee Walambe, Sheela Ramanna, and Ketan Kotecha. 2023. Domain adaptation: challenges, methods, datasets, and applications. *IEEE access*, 11:6973–7020.

Atnafu Lambebo Tonja, Bonaventure F. P. Dossou, Jessica Ojo, Jenalea Rajab, Fadel Thior, Eric Peter Wairagala, Anuoluwapo Aremu, Pelonomi Moiloa, Jade Abbott, Vukosi Marivate, and Benjamin Rosman. 2024. Inkubalm: A small language model for low-resource african languages.

Aminu Tukur, Kabir Umar, and Anas Sa'idu Muhammad. 2020. Parts-of-speech tagging of hausa-based texts using hidden markov model. *Dutse Journal of Pure and Applied Sciences (DUJOPAS)*, 6:303–313.

Pavanpankaj Vegi, Sivabhavani J, Biswajit Paul, Abhinav Mishra, Prashant Banjare, Prasanna K R, and Chitra Viswanathan. 2022. WebCrawl African : A multilingual parallel corpora for African languages. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1076–1089, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ludwig Wittgenstein. 1994. *Tractatus logico-philosophicus*. Edusp.

Lukas Wolf, Tiago Pimentel, Evelina Fedorenko, Ryan Cotterell, Alex Warstadt, Ethan Wilcox, and Tamar Regev. 2023. Quantifying the redundancy between prosody and text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9765–9784, Singapore. Association for Computational Linguistics.

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. Bloom: A 176b-parameter open-access multilingual language model.

S.A. Yakasai. 2025. Tauraruwa harshen hausa jiya da yau: Kalubale da madosa. *Tauraruwa Journal of Hausa Studies*, 1(1):1–9.

Yizhe Yang, Huashan Sun, Jiawei Li, Runheng Liu, Yinghao Li, Yuhang Liu, Yang Gao, and Heyan Huang. 2024. Mindllm: Lightweight large language model pre-training, evaluation and domain application. *AI Open*, 5:1 – 26.

Dong Yu and Lin Deng. 2016. *Automatic speech recognition*, volume 1. Springer.

Aliyu Yusuf, Aliza Sarlan, Kamaluddeen Usman Danyaro, and Abdullahi Sani BA Rahman. 2023. Finetuning multilingual transformers for hausa-english sentiment analysis. In *2023 13th International Conference on Information Technology in Asia (CITA)*, pages 13–18. IEEE.

Aliyu Yusuf, Aliza Sarlan, Kamaluddeen Usman Danyaro, Abdullahi Sani BA Rahman, and Mujaheed Abdullahi. 2024. Sentiment analysis in low-resource settings: A comprehensive review of approaches, languages, and data sources. *IEEE Access*.

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 (2279-0764)*, 10(4).

Abubakar Yakubu Zandam, Fatima Adam Muhammad, and Isa Inuwa-Dutse. 2023. Online threats detection in hausa language. In *4th Workshop on African Natural Language Processing*.

Wanru Zhao, Yihong Chen, Royson Lee, Xinchi Qiu, Yan Gao, Hongxiang Fan, and Nicholas D. Lane. 2024. Breaking physical and linguistic borders: Multilingual federated prompt tuning for low-resource languages.

Linan Zhu, Zhechao Zhu, Chenwei Zhang, Yifei Xu, and Xiangjie Kong. 2023. Multimodal sentiment analysis based on fusion methods: A survey. *Information Fusion*, 95:306–325.

6 Appendex

Table 1: Publicly available Hausa datasets

SN	Source	Domain	Task	Size	Repository
1	(Muhammad et al., 2022)	Tweets	Sentiment Analysis	30k	https://github.com/hausanlp/NaijaSenti/blob/main/README.md
2	Rakhmanov and Schlippe (2022a)	Teachers' evaluation	Sentiment Analysis	40k	https://github.com/MrLachin/HESAC
3	(Aliyu et al., 2022)	Tweets	Hate speech detection	6k	https://github.com/hausanlp/HERDPhobia
3	Adelani et al. (2023)	News	Topic classification	3k	https://github.com/masakhane-io/masakhane-news
4	(Inuwa-Dutse, 2023)	Tweets/News	Machine translation, raw texts		https://github.com/ijdutse/hausa-corpus/tree/master
5	(Dione et al., 2023)	News	POS tagging	1,504 sents.	https://github.com/masakhane-io/masakhane-pos/tree/main/data/hau
6	(Bichi et al., 2023)	News	Summarization	113 articles	https://journals.plos.org/plosone/article/file?type=supplementary&id=10.1371/journal.pone.0285376.s001
7	(Ogundepo et al., 2023)	Wikipedia	Question Answering	1171	https://github.com/masakhane-io/afriqa
8	(Adelani et al., 2021, 2022c)	NER	News	2,720 & 8,165	https://github.com/masakhane-io/masakhane-ner/
9	Adelani et al. (2022a)	Machine Translation	News		https://github.com/masakhane-io/lafand-mt/tree/main
10	(Akhbardeh et al., 2021)	Machine Translation	News & Religious	Numerous	https://data.statmt.org/wmt21/translation-task/
11	(Goyal et al., 2022)	Machine Translation	Wikimedia	∼2000	https://github.com/openlanguagedata/flores
12	(Vegi et al., 2022)	Machine Translation	Web Crawl		https://github.com/pavanpankaj/Web-Crawl-African?tab=readme-ov-file
13	(Sani et al., 2025a)	News	Text Classification	5172	https://github.com/TheBangis/hausa_corpus