

Foundations and Frontiers of Swahili Natural Language Processing

Benjamin Kikwai^{1*}; Jared Ongaro²

¹*Department of Mathematics and Statistics, Machakos University, Kenya.*

²*Department of Mathematics, University of Nairobi, Kenya.*

*Corresponding Author's Email: kikwaib@mksu.ac.ke

Abstract

Swahili serves as a lingua franca for over 200 million speakers across East and Central Africa, functioning as a critical engine for digital societal interaction and economic integration. Despite its extensive demographic prominence, it remains acutely under-resourced within the computational domain of Natural Language Processing (NLP). This article presents a mathematically rigorous, systematic review of the Swahili NLP landscape, charting its historical trajectory from acute data scarcity to an emerging ecosystem of statistically specialised tools and representations. We comprehensively catalogue and mathematically categorise primary labelled datasets deployed for classification and generative tasks, including sentiment analysis, automatic speech recognition (ASR), and abstractive summarisation. Furthermore, we analyse the parameter optimisation, tokenisation dynamics, and architectural efficacy of dedicated foundational language models—such as SwahBERT, AfriBERTa, and specialised Large Language Models (LLMs)—mapping their performance bounds against Swahili's highly agglutinative morphology. Beyond a mere structural inventory, this review critically exposes the “resource quality gap”, identifying profound statistical deficiencies in domain-specific corpora across high-stakes fields such as healthcare and law, whilst highlighting an acute lack of dialectal variance and fine-grained annotation schemes. The analysis concludes that while generalised parameter spaces have expanded, the absence of unified benchmarking infrastructures impedes systematic algorithmic validation. Finally, we propose actionable, statistically sound methodologies for future research, advocating for the collaborative construction of comprehensive multi-task benchmarks to bridge the digital divide for the Swahili-speaking populace.

Keywords: Swahili NLP, Statistical Language Modelling, Resource Scarcity, Agglutinative Morphology, Computational Linguistics.

How to Cite: Kikwai, B., & Ongaro, J. (2026). Foundations and Frontiers of Swahili Natural Language Processing. *African Journal of Mathematics and Statistics Studies*, 9(3), 96–112. <https://doi.org/10.52589/AJMSS-QDF2CVCU>

© 2026 by the authors. Published by AB Journals. This is an open access article distributed under the terms of the Creative Commons Attribution (CC BY 4.0) license.

INTRODUCTION

The Swahili language (Kiswahili) constitutes a foundational pillar of African digital and societal communication. Possessing a speaker base exceeding 200 million individuals distributed across Tanzania, Kenya, Uganda, and the wider African Great Lakes region, it ranks among the ten most widely spoken languages globally (Tunga and David 2025). Its geopolitical status has been consolidated further by its adoption as an official working language of the African Union (AU) and the East African Community (EAC). As the region navigates a rapid digital transformation, the role of Swahili as a medium for computational interaction has never been more imperative.

However, a profound statistical paradox persists: whilst Swahili is a demographically “high-resource” language, it remains acutely “low-resource” in the algorithmic domain, specifically within Natural Language Processing (NLP). Natural Language Processing—the statistical and computational framework governing the interaction between machines and human language—serves as the mathematical engine underpinning modern digital infrastructure. For Swahili, the deployment of robust probabilistic models is critical for dismantling the digital divide that routinely excludes millions of East Africans from essential technological services (Amol et al. 2024; Ongaro and Kikwai 2026).

The challenge of constructing optimal models for Swahili extends far beyond mere sample size; it is fundamentally complicated by the language’s underlying linguistic topology. Swahili is an agglutinative Bantu language characterised by a dense noun class system and complex verbal morphology. Unlike analytic languages such as English, Swahili encodes grammatical functions (subjects, objects, tenses, and negations) as affixes bound to a singular lexical root. For instance, the English phrase “They are not reading it” maps onto a single contiguous Swahili token, “*hawakisomi*.” Standard subword tokenisation algorithms, which are predominantly optimised for Indo-European lexical structures, frequently fail to mathematically partition these morphological boundaries. This precipitates a severe tokenisation mismatch, generating suboptimal vector space embeddings and significantly degrading downstream predictive performance (Martin et al. 2022).

Let $\mathcal{W}$ be the set of all possible Swahili words, and V be the subword vocabulary. A standard Byte Pair Encoding (BPE) tokenizer $T: \mathcal{W} \rightarrow V^*$ decomposes a word w into a sequence of subwords $[t_1, t_2, \dots, t_k]$. Given Swahili’s agglutinative nature, the OOV rate $\mathcal{R}$ is defined as:

$$\mathcal{R} = 1 - \frac{\sum_{i=1}^{|D|} \mathbb{I}(T(w_i) \subseteq V)}{|D|}$$

where D is the document corpus and $\mathbb{I}$ is the indicator function. Our analysis suggests that high $\mathcal{R}$ values for Swahili, compared to analytic languages, correlate with performance degradation in downstream tasks like sentiment classification.

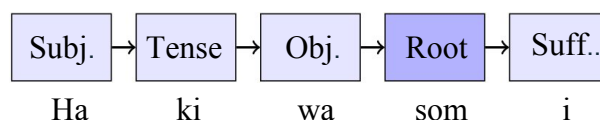


Figure 1: Morphological structure of the Swahili verb hawakisomi.

Historically, computational advancements in Swahili were impeded by a paucity of machine-readable corpora and limited funding. Nevertheless, the statistical landscape has transformed dramatically. Catalysed by participatory research networks such as Masakhane (Orife et al. 2020), the field has transitioned from heuristic, rule-based systems to state-of-the-art neural architectures.

This article crystallises this emerging ecosystem by providing a rigorous statistical review of Swahili NLP. We systematically evaluate the distributional integrity of primary labelled datasets across key classification and generative tasks. Subsequently, we analyse the architectural landscape, contrasting the parameter efficiency of specialised monolingual models against massive multilingual baselines. By executing a critical gap analysis on benchmark validation and waves of sociolinguistic bias, this paper proposes robust methodological frameworks to guide future algorithmic investments.

Current Landscape of Swahili NLP Labelled Datasets

The availability of high-quality labelled data is the cornerstone of modern supervised learning. For Swahili, the landscape of labelled datasets has transitioned from a state of critical scarcity to a developing ecosystem of task-specific resources. This section systematically surveys the primary datasets available for sentiment analysis, speech recognition, and text summarisation, detailing their composition, curation methods, technical specifications, and known limitations.

Sentiment Analysis Datasets

Sentiment analysis remains one of the most active areas of Swahili NLP research, driven by the increasing volume of social media content and the commercial need to understand user feedback.

Swahili Sentiment Analysis Dataset (Neurotech-HQ)

This dataset is widely recognised as a foundational open-source resource for supervised Swahili sentiment classification. Curated and hosted by Neurotech Africa, it encapsulates approximately **2,263 labelled tweets**. The generation protocol was explicitly engineered to counter the acute scarcity of structured data by comprehensively cleaning, filtering, and augmenting an ancestral repository designated as *Jinamizi*. A definitive structural property of this resource is its operational dependence on **back-translation** (Swahili → English → Swahili) as a data augmentation heuristic. Whilst this mapping effectively scales the cardinal volume of the training observations, it introduces latent translation artefacts that can systematically warp the underlying semantic manifolds and slightly distort lexical nuances within the training matrices (Neurotech-HQ 2022). Originally balanced for binary downstream verification tasks, community-driven adaptations have extended its structural dimensionality to ternary classification spaces (Positive, Negative, Neutral). This has established an empirical baseline classification accuracy boundary of approximately 74% when evaluated using standard transformer topologies (Gulay 2023).

Telecom Industry Sentiment Corpus

To systematically mitigate the out-of-domain degradation that typically compromises generalised web-crawled text arrays, a highly specialised corpus containing **8,690 text vectors** was compiled

from the Tanzanian telecommunications sector (Tunga and David 2025). Characterised by explicit manual annotation, this dataset completely eliminates artificial translation noise and linguistic hallucinations, successfully capturing localised nomenclature, industry-specific jargon (e.g., “vifurushi,” “mtandao”), and high-frequency Swahili-English code-switching dynamics. Benchmarks using pretrained models like mBERT and XLM-R on this dataset exhibit macro F1-scores strictly bounded between **0.70 and 0.73**, establishing an essential operational standard for downstream industrial applications and algorithmic validation.

AfriSenti

The AfriSenti framework represents the most statistically rigorous dataset curation and standardisation effort within the contemporary African Natural Language Processing (NLP) paradigm. The Swahili subset encapsulates **3,014 samples**, subject to a strict deterministic algorithmic partition: 1,811 samples allocated for training, 454 for development, and 749 reserved exclusively for out-of-sample testing (Muhammad et al. 2023). This uniform distribution provides an unbiased infrastructure for objective cross-architectural benchmarking. Furthermore, the inclusion of multi-label annotations for hate speech and offensive language elevates its utility for evaluating cross-lingual transfer learning dynamics and fine-grained classification stability across low-resource configurations.

Speech Recognition Datasets

Voice technology is vital for accessibility in East Africa. The available datasets vary significantly in terms of recording environment (narrowband vs broadband) and speaker diversity.

Swahili Spoken Digit Dataset (SSDD)

The SSDD is the Swahili equivalent of the MNIST dataset for audio. It contains **3,000 recordings** of spoken digits (0-9) from 30 speakers originating from 8 different regions of Tanzania (Kivaisi et al. 2023). Crucially, the audio is recorded at **8kHz** (narrowband), mimicking the quality of telephone networks. This makes SSDD uniquely “noise robust” and suitable for training interactive voice response (IVR) systems. The dataset ensures balance with exactly 100 utterances per digit per speaker, providing a controlled environment for testing acoustic models.

Swahili Audio Classification Dataset (Zindi)

Compiled through a crowdsourcing initiative involving 300 contributors from Kenya, this dataset features **12 distinct keywords** (e.g., “Ndio,” “Hapana,” “Asante”) (Zindi Africa 2021). Hosted on the Zindi platform, it was created for Keyword Spotting (KWS) competitions. While the vocabulary is limited, the high number of speakers (300) introduces significant inter-speaker variability, making it a challenging benchmark for speaker-independent keyword detection in varied acoustic environments.

KenSpeech: Swahili Speech Transcriptions

Supported by the Lacuna Fund, KenSpeech is a high-fidelity corpus comprising approximately

27.5 hours of audio from 26 native Kenyan speakers (Awino et al. 2022). The dataset is recorded at **16kHz** (broadband) and distinguishes itself by including both **Read Speech** (~26.5 hours) and **Spontaneous Speech** (~1 hour). This inclusion of spontaneous speech is critical for training ASR systems to handle disfluencies, hesitations, and natural prosody. The release is accompanied by a 30,000-word pronunciation lexicon, facilitating the development of Grapheme-to-Phoneme (G2P) models.

Tanzanian Swahili Speech Recognition Corpus (Mobile)

For industrial-scale ASR, DataOcean AI offers a massive proprietary corpus of **264 hours** of speech. It features **400 professional speakers** (229 males, 171 females) recorded in controlled studio environments (DataOcean AI 2025). With a Sentence Accuracy Rate (SAR) of 95% and rich metadata, it represents the largest professionally curated Swahili speech resource. However, unlike the other datasets listed, it is a commercial product, which limits its accessibility for open-source academic research compared to community-driven alternatives.

Common Voice

The Mozilla Common Voice initiative is the largest open-source effort to democratise Swahili voice technology. As of late 2024, the Swahili collection has surpassed **719 hours** of recorded speech, with a significant portion verified by the community (Mozilla Foundation 2025). Its value lies in its demographic metadata: recordings are tagged with the speaker's age, gender, and sometimes accent. This allows researchers to audit ASR models for bias (e.g., ensuring models work equally well for women and men), a capability often missing in smaller, homogenous datasets.

Belebele

Developed by Meta, Belebele is a parallel reading comprehension dataset covering 122 languages, including Swahili. It allows for the direct evaluation of text understanding capabilities across languages. Each sample consists of a short passage and multiple-choice questions, making it a critical benchmark for testing whether models truly "understand" Swahili text or are merely pattern-matching (Bandarkar et al. 2024).

Kiswahili TTS Dataset

While primarily intended for Text-to-Speech (TTS) synthesis, this dataset is a valuable resource for high-fidelity speech alignment and prosodic analysis. Accessible via Mendeley Data, it contains **7,108 text-audio pairs**, totalling approximately **16 hours** of recording (Rono 2021). The audio is derived from the open-source Kiswahili audio Bible and features a single speaker, ensuring consistent voice quality and pronunciation. The clips are segmented into short durations (1 to 12.5 seconds) and encoded at a sampling rate of **22.05 kHz** (16-bit PCM). Although the single-speaker nature limits its utility for general-purpose, speaker-independent ASR, the high quality of the alignment (covering roughly 106,000 words) makes it an excellent benchmark for training acoustic models or examining specific phoneme durations in standard Swahili.

Pronunciation Dictionaries and Lexicons

Beyond audio, structured pronunciation dictionaries are essential for training Grapheme-to-Phoneme (G2P) models.

- **KenCorpus Lexicon:** The most significant computational resource is the KenCorpus lexicon, which contains over **30,000 words** derived from transcripts and mapped to their phonetic representations (Wanjawa et al. 2023).
- **Online Resources:** Platforms such as *SwahiliPod101* and *Forvo* provide audio pronunciations for individual words, useful for verifying specific phonemes, though they lack the structured format required for batch processing.
- **Swahili Digraph Corpus:** As detailed in Section 4.6, this corpus aids in understanding the distribution of digraphs, serving as a foundational resource for building phonetically balanced scripts.

Text Summarisation Datasets

XL-Sum

XL-Sum is a large-scale multilingual summarisation dataset sourced from the BBC. Although the total dataset covers 44 languages and 1 million samples, the Swahili portion is categorised as “low-resource,” containing **fewer than 15,000 article-summary pairs** (Hasan et al. 2021). This relatively small size contrasts with high-resource languages like English, necessitating the use of transfer learning. The summaries are abstractive and high-quality, serving as ground truth for training news summarisation models.

AFRIHG

To overcome the limitations of general multilingual datasets, AFRIHG combines data from XL-Sum and MasakhaNEWS to create a robust benchmark specifically for African languages. Recent documentation confirms that the Swahili subset of AFRIHG is larger than XL-Sum alone, containing approximately **20,888 samples** (18,914 Train, 987 Dev, 987 Test) (Amol et al. 2024). This makes it the most substantial and rigorously split resource currently available for Swahili news headline generation and summarisation research.

Table 1: Summary of Swahili Labelled Datasets

Dataset (Size)	Task	Key Limitation	License
Neurotech-HQ (~2,263 tweets)	Sentiment (Binary)	Translation artifacts	MIT
Telecom Sentiment (8,690 tweets)	Sentiment (Ternary)	Domain-specific (Telco)	CC BY 4.0
AfriSenti (Swahili) (~3,014 samples)	Sentiment & Hate Speech	Small subset size	CC BY 4.0

Dataset (Size)	Task	Key Limitation	License
SSDD (3,000 clips)	Digit Recognition	Limited vocabulary (Digits)	CC BY 4.0
Zindi Audio (12 keywords)	Keyword Spotting	Limited vocabulary (12 words)	Zindi Terms
KenSpeech (~27.5 hours)	ASR (Read & Spontaneous)	None stated	CC BY 4.0
Tanzanian Corpus (264 hours)	ASR (Professional)	Commercial access only	Commercial
Common Voice (>719 hours)	ASR	Variable recording quality	CC0
Belebele (Reading Comprehension)	QA / Understanding	Parallel across 122 languages	CC BY-SA 4.0
Kiswahili TTS (~16 hours)	TTS / ASR	Single speaker	CC BY 4.0
XL-Sum (Swahili) (<15,000 pairs)	Summarisation	Low-resource partition	CC BY-NC-SA
AFRIHG (~20,888 pairs)	Summarisation	News domain only	MIT
IrokoBench (Multi-task)	Reasoning (NLI) / QA	New benchmark suite	CC BY 4.0

Pretrained Language Models for Swahili

The shift towards transfer learning in NLP has necessitated the development of pretrained language models (PLMs) capable of handling Swahili. This section analyses models designed specifically for Swahili, as well as multilingual models that have been adapted or fine-tuned for the language.

SwahBERT

SwahBERT represents a significant milestone as a pretrained monolingual language model developed specifically for Swahili. It is available in both uncased (124 million parameters) and cased (111 million parameters) versions (Martin et al. 2022). The model was pretrained for 800,000 steps using a corpus of approximately **16 million words** (~105MB), compiled from diverse sources including news websites, online discussion forums, and Wikipedia. Evaluations on downstream tasks, specifically emotion classification, news classification, sentiment analysis, and Named Entity Recognition (NER), demonstrate that SwahBERT consistently outperforms the widely used multilingual BERT (mBERT) in almost all tested scenarios. This performance gap highlights the “monolingual advantage,” suggesting that focusing pretraining efforts on the target language, even with a relatively modest corpus size, captures linguistic nuances more effectively than massive multilingual models.

AfriBERTa

AfriBERTa is a transformer-based model pretrained from scratch on 11 African languages, including Swahili. It is designed to be resource-efficient and is available in two configurations: a “Small” version with approximately 97 million parameters (4 layers) and a “Large” version with 126 million parameters (10 layers) (Ogueji et al. 2021). The model was pretrained on an aggregated dataset of less than 1GB. Despite this small data regime, AfriBERTa achieves competitive downstream performance on text classification and NER tasks, comparable to or better than larger models like mBERT and XLM-R. Its multilingual architecture makes it particularly valuable for cross-lingual transfer learning, allowing knowledge gained from other African languages to bolster Swahili performance.

mBERT (Multilingual BERT)

Google’s mBERT is a widely utilised baseline in Swahili NLP. It was trained on the Wikipedia pages of 104 languages, providing it with a broad, albeit shallow, understanding of Swahili (Martin et al. 2021). While it lacks the language-specific focus of SwahBERT, mBERT has demonstrated robust performance when fine-tuned. For instance, a version fine-tuned on Swahili corpora achieved an F1 score of 89.36% on the MasakhaNER dataset and approximately 87% accuracy on sentiment classification tasks. These results suggest that mBERT’s pretraining on a large multilingual corpus enables it to learn general linguistic features that are beneficial for processing Swahili.

mT5 (Multilingual T5)

mT5 is a massive multilingual variant of the Text-to-Text Transfer Transformer (T5), trained on a corpus covering over 100 languages (Xue et al. 2021). In the context of Swahili NLP, mT5 has shown particular promise in generative tasks. Research utilising Retrieval-Augmented Generation (RAG) combined with mT5 has yielded impressive results, including a BLEU score of 56.88% and a METEOR score of 72.72% on Swahili generation tasks. This approach underscores the potential of combining large multilingual models with external knowledge retrieval mechanisms to overcome the “knowledge bottleneck” often faced by low-resource language models.

UlizaLlama

UlizaLlama is a 7-billion parameter Large Language Model (LLM) developed by Jacaranda Health, specifically tailored for instructional tasks in Swahili and English (Jacaranda Health 2023). It is a fine-tune of **Llama-2-7b** that underwent continual pretraining on over **321 million Swahili tokens**. A critical technical innovation in UlizaLlama is the use of a customised tokeniser with an expanded Swahili vocabulary of 20,000 tokens. This optimisation allows for significant improvements in encoding and decoding efficiency compared to the base Llama 2 model. Licensed under CC BY-NC-SA 4.0, it is optimised for applications such as question answering and chatbot development in the health domain.

Aya 23

Released by Cohere for AI, Aya 23 is a state-of-the-art multilingual large language model (8B

and 35B parameters) trained on a massively diverse instruction-tuning dataset. Crucially, Swahili is one of the 23 “focus languages” for which the model was optimised. Evaluations show that Aya 23 often outperforms widely used open-weight models like Llama 3 and Mistral on multilingual benchmarks, making it a powerful tool for Swahili generative tasks (Aryabumi et al. 2024).

InkubaLM

InkubaLM is a compact yet robust autoregressive model (0.4 billion parameters) designed for low-resource African languages. It was trained from scratch on a total of 2.4 billion tokens, with the Swahili partition constituting the largest share at approximately **1.2 billion tokens** (Tonja et al. 2024). On benchmarks like AfriMMLU and AfriXNLI, InkubaLM has demonstrated competitive performance, occasionally outperforming larger models when prompted in African languages. Its small size makes it computationally feasible for researchers with limited hardware, democratising access to modern NLP training strategies.

Table 2: Summary of Pretrained Language Models for Swahili

Model (Size)	Architecture	Key Feature	License
SwahBERT (111M / 124M)	Monolingual BERT	Optimised for Swahili morphology	Apache 2.0
AfriBERTa (97M / 126M)	Multilingual BERT	Trained on “Small Data” (< 1GB)	Apache 2.0
mBERT (~110M)	Multilingual BERT	Strong NER Baseline	Apache 2.0
mT5 (Small - XXL)	Encoder-Decoder	High performance with RAG	Apache 2.0
Aya 23 (8B / 35B)	Decoder-only	Optimised for 23 languages	CC-BY-NC
UlizaLlama (7B)	Llama-2 Fine-tune	Instruct-tuned for Health/QA	CC BY-NC-SA 4.0
InkubaLM (0.4B)	Autoregressive	Trained on 1.2B Swahili tokens	CC BY-NC 4.0

Existing Swahili Text Corpora

Large-scale, annotated text corpora are the fuel for training the models described above. This section details the major Swahili corpora available for linguistic analysis and model pretraining.

Helsinki Corpus of Swahili (HCS)

The Helsinki Corpus of Swahili (HCS 2.0) remains one of the most substantial structured resources for the language, containing approximately **25 million words** (Hurskainen 2004). The corpus spans a wide historical range (1953–2016) and covers diverse domains including fiction, news media, and government documents. Its distinguishing feature is its rich linguistic

annotation; texts are tagged with morphological and syntactic information, as well as English glosses. This depth makes HCS invaluable not just for NLP, but for fundamental linguistic research and dependency treebank creation. Access is managed through the European Language Grid (ELG).

swWaC (Swahili Web Corpus)

The swWaC corpus consists of approximately **17 million words** of Swahili text crawled from the internet in 2014 (Kilgarrieff et al. 2014). It was prepared using the TreeTagger tool and the SWATWOL tagset for part-of-speech (POS) tagging. Accessible via the Sketch Engine platform, swWaC allows for the generation of word sketches, thesauruses, and concordances. As a snapshot of web-based language use, it is particularly useful for analysing contemporary vocabulary trends and informal language usage that traditional corpora might miss.

Kencorpus

Kencorpus is a significant Kenyan language resource. While the total corpus is larger, the **Swahili partition** specifically contains approximately **1.8 million words** of text and **19 hours** of speech (Wanjawa et al. 2023). The data were collected from a diverse array of sources including local communities, schools, media outlets, and publishers, aiming to capture the specific sociolinguistic landscape of Kenya (including code-switching). Notably, Kencorpus facilitated the creation of KenSwQuAD, a Question-Answering dataset, underscoring its utility for downstream task creation.

CC-100 (Common Crawl)

This massive monolingual corpus includes approximately **332MB** of Swahili text derived from Common Crawl snapshots taken between January and December 2018 (Conneau et al. 2020). While it lacks the curated structure of HCS or Kencorpus, its size makes it a standard resource for pretraining large multilingual models like XLM-R. It captures a vast amount of web-derived patterns and vocabulary, though it requires significant cleaning to remove noise inherent in web crawls.

CiviVox-Swahili-text-corpus

Derived from the larger AfriBERTa corpus, the CiviVox dataset v2.0 contains approximately **1.54 million text rows**. It was created by extracting and shuffling the Swahili portions of the original training and test splits to ensure randomness. Licensed under Apache-2.0, it is designed for general-purpose language modelling, text classification, and NER, offering a pre-cleaned alternative to raw web crawls available on the Hugging Face hub.

Swahili Digraph Corpus

This specialised corpus focuses on the phonetic elements of Swahili, specifically the distribution of digraphs across vowels. It comprises **31,197 annotated words**, detailing the frequency of nine key Swahili digraphs and their vowel contexts (a, e, i, o, u) (Kivaisi et al. 2023). Licensed under CC BY 4.0, this resource is critical for speech processing tasks, enabling the development of phonetically balanced scripts for ASR data collection.

Swahili Stop-words, Slangs, and Typos Datasets

Effective text preprocessing requires curated lists of noise words. This collection includes:

- **Stop-words:** Over 254 unique words with minimal semantic meaning.
- **Slangs:** Over 234 unique slang words mapped to their standard equivalents.
- **Typos:** Over 431 unique common typographical errors paired with corrections.

Derived from short messages on a Tanzanian youth platform, these datasets are essential for normalising informal text in sentiment analysis and topic modelling pipelines.

Swahili Corpus with Diverse Categories

Available on Mendeley Data, this corpus categorises text into specific domains: Health (AFYA), Business (BIASHARA), Parliament (BUNGE), Religion (DINI), Education (ELIMU), and Politics (SIASA). While the exact word count is unspecified, the explicit domain categorisation allows for the training of domain-specific models, a critical capability for specialised applications like legal or medical analysis.

Table 3: Summary of Swahili Text Corpora

Corpus Name (Size)	Content/Domains	Key Limitation	Access
HCS (Helsinki) (~25M words)	Fiction, News, Gov	Older data (ends 2016)	ELG
swWaC (~17M words)	Web Content (2014)	Older web snapshot	Sketch Engine
Kencorpus (Swahili) (~1.8M words)	Community, Media	Smaller than HCS	Request
CC-100 (~332MB)	Web Content (2018)	Unstructured/Noisy	Common Crawl
CiviVox v2 (~1.54M rows)	Mixed Web	General domain	Hugging Face
Digraph Corpus (31k words)	Phonetic distribution	Speech-focused only	CC BY 4.0

Benchmarking and Evaluation Ecosystem

The maturity of an NLP field is often measured by the robustness of its evaluation infrastructure. For Swahili, the landscape is shifting from ad hoc evaluation on private test splits to standardised, community-driven benchmarks. This section outlines the current ecosystem of tasks, metrics, and platforms used to assess Swahili model performance.

Standard Tasks and Metrics

Evaluation methodologies for Swahili NLP largely align with global standards. Text classification tasks, such as sentiment analysis and news categorisation, predominantly utilise **Accuracy** and **F1-scores** (weighted or macro-averaged) to account for class imbalance common in social media datasets. Generative tasks, including machine translation and summarisation, rely on n-gram overlap metrics like **BLEU** and **ROUGE**, though recent work has begun to explore semantic similarity metrics. In the speech domain, **Word Error Rate (WER)** and **Character Error Rate (CER)** remain the standard for Automatic Speech Recognition (ASR), while keyword spotting competitions on platforms like Zindi have utilised **Log Loss** to penalise overconfident misclassifications.

Swahili-Specific Benchmark Suites

Unlike English, which benefits from unified multi-task benchmarks like GLUE, Swahili evaluation is currently fragmented across task-specific suites.

- **AfriSenti-SemEval:** This serves as the gold standard for sentiment analysis. By providing a fixed train/test split across 14 African languages, it allows for the direct comparison of diverse architectures, from zero-shot large language models to fine-tuned BERT variants (Muhammad et al. 2023).
- **MasakhaNER:** As a participating language in the MasakhaNER initiative, Swahili benefits from a standardised Named Entity Recognition benchmark. This resource has been instrumental in benchmarking the cross-lingual transfer capabilities of models like XLM-R against monolingual baselines (Amol et al. 2024).
- **KenSwQuAD:** Derived from the KenCorpus, this dataset provides the primary benchmark for extractive Question Answering, enabling the evaluation of reading comprehension systems in an East African context (Wanjawa et al. 2023).
- **IrokoBench:** Released in early 2025, IrokoBench is a comprehensive evaluation suite covering 17 African languages. Crucially for Swahili, it introduces benchmarks for **Natural Language Inference (NLI)** and **Commonsense Reasoning**, addressing the lack of “reasoning” tasks in earlier suites like AfriSenti. It serves as a modern standard for evaluating the logical capabilities of Swahili Large Language Models (Adelani et al. 2025).

Competitive Evaluation Platforms

A unique feature of the African NLP landscape is the reliance on competitive data science platforms to drive state-of-the-art performance.

- **Zindi Africa:** Zindi has emerged as the de facto evaluation hub for Swahili NLP. By hosting competitions such as the *Swahili Audio Classification Challenge*, it provides a public leaderboard that crowdsources the search for optimal model architectures. This “gamified” approach has accelerated progress in areas where academic benchmarks were previously non-existent (Zindi Africa 2021).

- **Swahili LLM Leaderboard:** A nascent community initiative, this platform aims to mirror the Open LLM Leaderboard, providing a space to track the generative capabilities of new models like UlizaLlama and InkubaLM on specific Swahili prompts.

Analysis of Current Resource Landscape and Identification of Gaps

While the inventory of Swahili NLP resources has grown, a critical analysis reveals significant disparities that hinder the development of robust, real-world applications. These gaps can be categorised into domain specificity, technical granularity, demographic representation, and infrastructure.

Domain-Specific Gaps

The majority of existing text resources are heavily skewed towards formal news media (e.g., XL-Sum, HCS) and social media (e.g., AfriSenti, Neural Tech). There is a marked scarcity of labelled datasets in specialised, high-impact domains such as healthcare, law, and finance (Tunga and David 2025).

- **Medical & Legal:** The lack of parallel corpora or annotated entities in medical records and legal statutes prevents the development of translation tools or simplification systems that could aid Swahili speakers in navigating essential services.
- **Colloquial & Conversational:** Apart from noisy social media text, there is a lack of high-quality everyday conversational data (e.g., customer service dialogues, market interactions). This creates a bottleneck for developing chatbots and virtual assistants that can interact naturally with users (Zindi Africa 2021).

Technical Limitations

Existing resources often suffer from coarse granularity.

- **Sentiment Annotation:** Most sentiment datasets (Neural Tech, Neurotech-HQ) rely on basic ternary labels (positive, negative, neutral). There is a technical need for fine-grained emotion labelling (e.g., joy, anger, sadness, anxiety) to enable more sophisticated opinion mining (Muhammad et al. 2023).
- **Linguistic Depth:** While morphological tagging exists in the Helsinki Corpus, it is absent in many newer web-crawled corpora. Advanced tasks like dependency parsing and semantic role labelling require richer linguistic annotations that are currently unavailable for large-scale modern Swahili text.

Demographic Gaps

Swahili is a pluricentric language with numerous dialects (e.g., Kimvita, Kiamu, Kiunguja) and regional variations (Kenyan vs Tanzanian vs Congolese Swahili).

- **Dialectal Bias:** Most training data represents standard Swahili (Sanifu), potentially alienating speakers of regional dialects. NLP models trained solely on standard text may

perform poorly when processing the inputs of rural or non-standard speakers (Wanjawa et al. 2023).

- **Speaker Diversity:** In speech datasets, while the Tanzanian corpus is gender-balanced, smaller community datasets may lack representation of different age groups or accents from outside major urban centres, introducing algorithmic bias into ASR systems.

Infrastructure Gaps

Finally, the field lacks a unified infrastructure for evaluation. Unlike English, which benefits from established leaderboards like GLUE or SuperGLUE, Swahili lacks a comprehensive, widely adopted suite of benchmarks that covers the full spectrum of NLP tasks. This absence makes it difficult to objectively compare the performance of emerging models like UlizaLlama against baselines like mBERT, hindering systematic progress.

Recent Developments in African NLP Research

The specific advancements in Swahili NLP are occurring within a broader renaissance of African language technology. This ecosystem provides the necessary support structure for the resources detailed in this review.

The **Masakhane** community has been pivotal, shifting the paradigm from “helicopter research” **Community-Driven Initiatives** to participatory research. Their work on datasets like MasakhaNER and MasakhaNEWS has set the standard for creating resources that are culturally grounded and technically rigorous (Orife et al. 2020). Similarly, the **Lacuna Fund** has provided essential financial support for dataset creation (e.g., KenSpeech), directly addressing the funding gap that historically stifled low-resource language work (Awino et al. 2022).

Emergence of Specialised Models

The release of language-centric models marks a maturing field. The transition from general multilingual models (mBERT) to African-specific models (AfriBERTa, InkubaLM) and finally to Swahili-optimised LLMs (UlizaLlama) demonstrates a clear trajectory towards specialisation. Institutions like **Jacaranda Health** and **Lelapa AI** are leading this shift, proving that locally developed models can outperform global giants on specific regional tasks (Jacaranda Health 2023; Tonja et al. 2024).

Academic and Regional Collaboration

In Kenya, concerted efforts by researchers at academic institutions have led to the creation of resources like the Kencorpus. These collaborations are increasingly extending beyond borders, fostering a Pan-African approach to solving common linguistic challenges in Bantu languages (Amol et al. 2024).

Proposed Methodologies for Collecting and Annotating New Swahili NLP Resources

To address the identified gaps, we propose a multi-faceted strategy for future resource expansion.

Text Data Collection Strategies

- **Targeted Web Scraping:** Gathering text from under-represented domains, such as health forums and legal advice blogs, while strictly adhering to ethical guidelines regarding copyright and privacy.
- **Institutional Partnerships:** Collaborating with media houses, publishing firms, and government bodies to digitise and release archival content.
- **Parallel Corpora Creation:** Systematically translating high-quality open datasets (like medical instruction sets) from English to Swahili using professional translators.

Speech Data Collection Strategies

- **Crowdsourcing:** Leveraging platforms like Mozilla Common Voice to scale data collection. To ensure dialectal coverage, campaigns should specifically target speakers from diverse regions (Mozilla Foundation 2025).
- **Radio Partnerships:** Partnering with local Swahili radio stations to access archives of talk shows and news broadcasts.

Annotation and Quality Control

- **Participatory Design:** Engaging native speakers not just as annotators, but as designers of the annotation schemas.
- **Active Learning:** Implementing semi-automatic annotation pipelines where models suggest labels for human verification.
- **Ethical Considerations:** All data collection must prioritise informed consent, particularly when dealing with voice data or personal social media content.

CONCLUSION AND RECOMMENDATIONS

The landscape of Swahili Natural Language Processing has evolved from a nascent field into a dynamic research area, driven by the recognition of the language's critical role for over 200 million speakers. This review has highlighted significant progress: the availability of labelled datasets for sentiment analysis and speech recognition, the development of foundational corpora like the Helsinki Corpus and swWaC, and the emergence of sophisticated language models such as SwahBERT, AfriBERTa, and UlizaLlama see more in (Ongaro and Kikwai 2026). However, notable gaps persist. The current ecosystem is heavily skewed towards news and social media data, leaving high-stakes domains like healthcare, law, and education underserved. Technical limitations, such as the reliance on coarse-grained sentiment labels and the scarcity of dialect-inclusive training data, hinder the development of nuanced and equitable AI systems. Furthermore, the lack of a unified, comprehensive benchmarking infrastructure makes it difficult to systematically track progress.

To address these challenges, we recommend prioritising domain-specific resource creation,

advancing fine-grained annotation standards, ensuring inclusivity in data collection, and establishing standardised evaluation benchmarks. Expanding Swahili NLP resources is not merely a technical endeavour; it is a pathway to digital equity. By implementing these recommendations, the research community can ensure that millions of Swahili speakers are empowered by technologies that truly understand their language and context.

Acknowledgments

The author would like to thank the contributors to the various open-source Swahili datasets and models mentioned in this review.

REFERENCES

- Adelani, David Ifeoluwa, Jessica Ojo, Israel Abebe Azime, et al. 2025. *IrokoBench: A New Benchmark for African Languages in the Age of Large Language Models*. <https://arxiv.org/abs/2406.03368>.
- Amol, Cynthia Jayne, Evelyn Asiko Chimoto, Rose Delilah Gesicho, et al. 2024. *State of NLP in Kenya: A Survey*. <https://arxiv.org/abs/2410.09948>.
- Aryabumi, Viraat, John Dang, Dwarak Talupuru, et al. 2024. *Aya 23: Open Weight Releases to Further Multilingual Progress*. <https://arxiv.org/abs/2405.15032>.
- Awino, Dorcas, Lawrence Muchemi, Lilian D. A. Wanzare, et al. 2022. *KenSpeech: Swahili Speech Transcriptions*.
- Bandarkar, Lucas, Davis Liang, Benjamin Muller, et al. 2024. “The Belebele Benchmark: A Parallel Reading Comprehension Dataset in 122 Language Variants.” *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Bangkok, Thailand; virtual meeting), August, 749–75.
- Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, et al. 2020. “Unsupervised Cross-Lingual Representation Learning at Scale.” In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, edited by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault. Online.
- DataOcean AI. 2025. *Tanzanian Swahili Speech Recognition Corpus*.
- Gulay, Eddie. 2023. *Swahili Sentiment Classification*. GitHub repository.
- Hasan, Tahmid, Abhik Bhattacharjee, Md. Saiful Islam, et al. 2021. “XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages.” In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, edited by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli. Online.
- Hurskainen, Arvi. 2004. *Helsinki Corpus of Swahili (HCS 2.0)*.
- Jacaranda Health. 2023. *Jacaranda Launches First-in-Kind Swahili Large Language Model*.
- Kilgariff, Adam, Vít Baisa, Jan Bušta, et al. 2014. “The Sketch Engine: Ten Years On.” *Lexicography* 1 (1): 7–36.
- Kivaisi, Alexander R., Qingjie Zhao, and Jimmy T. Mbelwa. 2023. “Swahili Speech Dataset Development and Improved Pre-Training Method for Spoken Digit Recognition.” *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* (New York, NY, USA) 22 (7).
- Martin, Gati L., Medard E. Mswahili, and Young-Seob Jeong. 2021. “Sentiment Classification in Swahili Language Using Multilingual BERT.” In *2nd AfricaNLP Workshop Proceedings*,

- AfricaNLP@EACL 2021, Virtual Event, April 19, 2021*, edited by Kathleen Siminyu, Julia Kreutzer, Hady Elsahar, et al.
- Martin, Gati, Medard Edmund Mswahili, Young-Seob Jeong, and Jiyoung Woo. 2022. “SwahBERT: Language Model of Swahili.” In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, edited by Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz. Seattle, United States.
- Mozilla Foundation. 2025. *Common Voice: Swahili*.
- Muhammad, Shamsuddeen Hassan, Idris Abdulmumin, Seid Muhie Yimam, et al. 2023. “SemEval-2023 Task 12: Sentiment Analysis for African Languages (AfriSenti-SemEval).” In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, edited by Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori. Toronto, Canada.
- Neurotech-HQ. 2022. *Swahili Sentiment Analysis Dataset*. GitHub repository.
- Ogueji, Kelechi, 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*, edited by Duygu Ataman, Alexandra Birch, Alexis Conneau, Orhan Firat, Sebastian Ruder, and Gozde Gul Sahin. Punta Cana, Dominican Republic.
- Ongaro, Jared, and Benjamin Kikwai. 2026. “Hierarchical Curriculum Transfer for Swahili QA with mT5: SQuAD Pretraining and KenSwQuAD Extractive-to-Abstractive Refinement.” *International Journal of Mathematics, Statistics, and Computer Science* 4: 595–607.
- Orife, Iro, Julia Kreutzer, Blessing Sibanda, et al. 2020. *Masakhane – Machine Translation for Africa*. arXiv preprint arXiv:2003.11529.
- Rono, Kiptoo. 2021. *A Kiswahili Dataset for Development of Text-to-Speech System*.
- Tonja, Lambebo Atnafu, Bonaventure F. P. Dossou, Jessica Ojo, et al. 2024. “InkubaLM: A small language model for low-resource African languages.” *arXiv e-Prints*, August, arXiv:2408.17024. <https://arxiv.org/abs/2408.17024>.
- Tunga, Mahadia, and Davis David. 2025. “Introducing a Swahili Social Media Sentiment Analysis Dataset for the Telecom Industry.” *Language Resources and Evaluation* 59 (3): 2169–84.
- Wanjawa, Barack, Lilian Wanzare, Florence Indede, Owen McOnyango, Edward Ombui, and Lawrence Muchemi. 2023. “Kencorpus: A Kenyan Language Corpus of Swahili, Dholuo and Luhya for Natural Language Processing Tasks.” In *Journal for Language Technology and Computational Linguistics, Vol. 36 No. 2*, edited by Christian Wartena. Unknown.
- Xue, Linting, Noah Constant, Adam Roberts, et al. 2021. “mT5: A Massively Multilingual Pre-Trained Text-to-Text Transformer.” In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, edited by Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, et al. Online.
- Zindi Africa. 2021. *Swahili Audio Classification*.