Applied AI Letters

WILEY

LETTER OPEN ACCESS

Building Text and Speech Benchmark Datasets and Models for Low-Resourced East African Languages: Experiences and Lessons

Joyce Nakatumba-Nabende[1] | Claire Babirye[2] | Peter Nabende[3] | Jeremy Francis Tusubira[2] | Jonathan Mukiibi[2] | Eric Peter Wairagala[2] | Chodrine Mutebi[2] | Tobius Saul Bateesa[2] | Alvin Nahabwe[2] | Hewitt Tusiime[2] | Andrew Katumba[4]

[1]Department of Computer Science, Makerere University, Kampala, Uganda | [2]Makerere Artificial Intelligence Lab, Makerere University, Kampala, Uganda | [3]Department of Information Systems, Makerere University, Kampala, Uganda | [4]Department of Electrical and Computer Engineering, Makerere University, Kampala, Uganda

Correspondence: Joyce Nakatumba-Nabende (joyce.nabende@mak.ac.ug)

Received: 30 March 2023 | **Revised:** 21 December 2023 | **Accepted:** 22 February 2024

Funding: This work was carried out with support from Lacuna Fund—No. 1937.70 2021, an initiative co-founded by The Rockefeller Foundation, Google.org, and Canada's International Development Research Centre, Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ), and Mozilla.

Keywords: automatic speech recognition | low-resourced language | machine translation | speech dataset | text dataset | topic modeling

ABSTRACT

Africa has over 2000 languages; however, those languages are not well represented in the existing natural language processing ecosystem. African languages lack essential digital resources to effectively engage in advancing language technologies. There is a need to generate high-quality natural language processing resources for low-resourced African languages. Obtaining high-quality speech and text data is expensive and tedious because it can involve manual sourcing and verification of data sources. This paper discusses the process taken to curate and annotate text and speech datasets for five East African languages: Luganda, Runyankore-Rukiga, Acholi, Lumasaba, and Swahili. We also present results obtained from baseline models for machine translation, topic modeling and classification, sentiment classification, and automatic speech recognition tasks. Finally, we discuss the experiences, challenges, and lessons learned in creating the text and speech datasets.

1 | Introduction

Uganda is a multilingual country with over 41 spoken and indigenous languages [1]. However, these languages are scarcely represented electronically. Like many African languages, Ugandan languages face the challenges of not having easily accessible electronic resources for building downstream natural language processing (NLP) datasets, research, and tools [2, 3] despite the substantial number of speakers of these languages. The text and speech resources are limited because collecting, building, and processing these resources are costly and time-consuming. Text data resources are also hard to discover for the low-resourced languages [4]. The lack of text data resources for these languages has affected the advancement of research in various NLP applications. While some previously low-resourced African languages currently have NLP resources, many other indigenous African languages are either not resourced and do not have sufficient NLP resources, including datasets or very low-resourced. This is the case for the majority of the 41 indigenous languages. It has been noted in [5] that NLP tasks that benefit the African context, such as machine translation (MT), text classification, topic modeling, speech recognition, and named entity recognition (NER), require resources and knowledge that the native language speakers can provide. These native speakers should collaboratively be involved in developing the NLP datasets and models [5, 6]. Therefore, there is a need to generate high-quality NLP

resources for local languages spoken in African countries [5, 7]. However, obtaining such high-quality data is expensive and tedious because it can involve manual sourcing and verification of data sources.

Automatic speech recognition (ASR) has recently attained state-of-the-art performance levels for numerous Western and Asian languages such as English [8], French [9, 10], Chinese [11], and Japanese [12]. This success can be attributed to the abundance of high-quality speech resources. However, African languages have yet to benefit from widespread ASR applications. This discrepancy is primarily a result of the insufficient availability of speech resources for the majority of African languages [13].

The main official languages in Uganda are English and Swahili, which are not the native languages spoken in Uganda. Although English is the official language in Uganda, we do not include it in this paper as it is a highly resourced language and its resources are available electronically. The focus of this paper is to create text and speech datasets for four major Ugandan languages: Luganda, Runyankole-Rukiga, Lumasaba, and Acholi and the major spoken language in East Africa, which is Swahili. We create a diverse text dataset in four Ugandan languages and one East African language that consists of both formal and informal types of languages from the native language speakers in East Africa. The main contribution of this work is to aid in creating diverse text and speech datasets for East African languages.

a. We make available monolingual corpus for four Ugandan languages—Luganda, Runyankole-Rukiga, Lumasaba, and Acholi—and one East African language: Swahili.

b. We make available parallel corpus from English to five East African languages: Luganda, Runyankole-Rukiga, Lumasaba, Acholi, and Swahili.

c. We create large speech corpus for Luganda and Swahili on the Mozilla's Common Voice (CV) platform.

d. We leverage the text and speech datasets to build baseline models for specific downstream tasks: MT, topic modeling and classification, sentiment analysis, and ASR tasks.

e. We provide insights into experiences and lessons learned on building text and speech datasets for low-resourced languages.

The rest of the paper is organized as follows: Section 2 discusses the related work. Section 3 describes the languages created in this paper. Sections 4 and 5 describe the process of creating the text and speech corpus, respectively. Section 6 describes the tools used to collect the text and speech dataset. Section 7 provides a description of the quality assurance steps taken with the data preparation. Section 8 discusses the benchmark results from MT, topic modeling and classification, and ASR experiments based on the text and speech datasets. Section 9 discusses the key challenges and lessons learned, and finally, Section 10 concludes the paper.

2 | Related Work

Efforts have been made in NLP to collect data, build NLP research communities, and advance NLP research efforts on the African continent. Recently, the Lacuna fund [14] was created to support the researchers in the African continent to create openly accessible text and speech resources for building NLP technologies in diverse languages across Africa. The fund has fueled and accelerated the creation of language datasets for diverse low-resourced languages. In this section, we review the existing approaches to create text and speech datasets for various low-resourced languages.

2.1 | Text Datasets

The most publicly available text corpus range from MT [15–17], sentiment analysis [18–21], and NER [5, 22, 23]. The most common being the MT or parallel translated datasets. Masakhane [24] is one of the communities that has enhanced NLP research in Africa through a participatory approach to community building [25]. Masakhane has created and publicly released several MT datasets and baseline models.[1] The most notable is MAFAND-MT, a news domain parallel corpus for 16 African languages [26]. Other efforts under Masakhane have been translating Edoid languages, MT from Fon to French [27], MT for Nigerian Pidgin [28], and many others. The Masakhane datasets are created and curated by various volunteers from various parts of Africa. The teams of volunteers are usually natives, language experts from universities, researchers, and other community volunteers who contribute well-curated translations. Masakhane has also worked on other NER datasets such as MasakhaNER [5] and MasakhaNER 2.0 [29]. Sunbird AI [30] has also recently started efforts to create open datasets for African languages. The Sunbird African Language Technology (SALT) dataset is a parallel MT corpus for five low-resourced Ugandan languages [31].

The dataset was created by translating English sentences sourced from various domains. The translators were professional translators, language experts, teachers, and tutors of the languages. The English sentences extracted from various sources were used as prompts for creating new simplified English sentences translated by the translators. The Artificial Intelligence for Development (AI4D) program [32] has also supported the collection and creation of MT datasets like the Menyo20K [15], a multidomain English to Yoruba corpus and an English to Luganda parallel text corpus [33]. The English to Luganda parallel text corpus was created by the team at Makerere Artificial Intelligence Lab, which was the starting point for the data collection efforts reported in this paper. Recently, two multilingual government-themed corpus in various South African languages known as “The Vuk'uzenzele South African Multilingual Corpus” were created by gathering the South African government newspapers translated into all 11 South African official languages [34]. The dataset was created as a collection of magazine PDF editions manually cleaned and reviewed. Work has been carried out to create the Kencorpus dataset [35] which is a Kenyan language corpus for Swahili, Dholuo, and Luhya languages. The Kencorpus contains a subset of Dholuo to Swahili translations and Luhya to Swahili translations. Other text datasets included in the Kencorpus are question–answering and POS tagging datasets.

Efforts have also been made to create and curate a news headline dataset for two low-resourced languages in South Africa: Setswana and Sepedi [3]. The dataset includes 219 news headlines in Setswana and 491 news headlines in Sepedi. The news

headlines were categorized into several topics in the dataset. More work in NLP for African languages has been done in sentiment analysis. The AfriSenti [20] corpus consists of 14 sentiment datasets of 110,000+ tweets in 14 African languages (Amharic, Algerian Arabic, Hausa, Igbo, Kinyarwanda, Moroccan Arabic, Mozambican Portuguese, Nigerian Pidgin, Oromo, Swahili, Tigrinya, Twi, Xitsonga, and Yorùbá) from four language families annotated by native speakers. The tweets were collected through the Twitter Academic API using stop-words, sentiment lexicons, and a language detection tool. The tweets were then anonymized, preprocessed, and annotated by three native speakers. The text datasets have been created for several downstream tasks, such as MT, sentiment analysis, named entity analysis, and question–answering, as discussed and documented in the papers.

2.2 | Speech Datasets

Mozilla's CV project[2] through its web platform has led tremendous efforts toward the creation of publicly available datasets. Various African languages have been added to the CV platform. Kinyarwanda was the first African language launched on the CV platform. The Kinyarwanda CV dataset currently has over 2300 hours of recorded voice and 1900 hours of validated voice. There have been several additions of African languages on the CV platform, including Luganda, Swahili, Igbo, and Yoruba. The approach taken with voice contributions has been community mobilizations of voice contributors across the several countries where the languages are spoken. The Kencorpus [35], which contains 177 hours of speech data for Swahili, Dholuo, and Luhya languages, was collected using voice recorders and the rest was obtained through collaborating with media houses.

While the CV datasets are of a read speech type, other efforts exist to create speech datasets for other domains. The ZA-Gov Multilingual South African corpus is a dataset of South African government speeches [34]. The dataset was created from the speeches following cabinet meetings of the South African government. The dataset contains topics from various domains such as energy, labor, service delivery, crime, COVID-19, and cabinet decisions. The data were scraped from the South African government website, where all cabinet statements and translations are posted. The Makerere AI Lab also released the "Makerere Radio Speech Corpus", a Luganda radio corpus for ASR used to build a COVID-19 speech-to-text (STT) model [36]. The dataset was collected from online radio streams, cleaned, and transcribed by professional translators.

The AI4D fellowship also supported the creation of an agricultural-specific speech keyword dataset [37] that contains 5290 keyword utterances of specific keywords in English and Luganda. While collecting the keywords dataset, different language coordinators contacted community members. They encouraged the community members to read the keywords on the web platform as they recorded themselves reading out aloud. Language experts then validated the recorded utterances to create a well-curated dataset. The Luganda keyword utterances dataset was used in a machine learning competition hosted on the Zindi platform.[3] The authors in [38] developed a mobile application for collecting parallel speech data for under-resourced languages for three languages: Mboshi, Myene, and Basaa. The increasing availability of African datasets has led to the development of platforms like Lanfrica,[4] which catalogs and links African language resources to mitigate the difficulty encountered in the discovery of African language datasets. The work presented in this paper builds upon this existing work, and we present an approach taken to create speech and text datasets for low-resourced East African languages.

3 | Language Characteristics

In this work, we focused on developing text and speech language resources for five sub-Saharan African languages with varying numbers of speakers between 1.5M and 100M spoken in around 10 countries in the Eastern, Central, and Southern regions of Africa. The selected languages cover two language families. Three of the languages belong to the Niger-Congo language family, and one language belongs to each of the Western-Nilotic family. Table 1 provides an overview of the languages in our corpus.

Although many languages belong to the Niger-Congo-Bantu language families, they have different linguistic characteristics as described below.

3.1 | Luganda

(lug) has a subject–verb–object word order and is a tonal language [40]. Luganda's alphabet is composed of 24 letters that include: 18 consonants: b, p, v, f, m, d, t, l, r, n, z, s, j, c, g, k, ny, ŋ, 5 vowel sounds represented in the five alphabetical symbols (a, e, i, o, u) and 2 semi-vowels (w, y). The special consonant ŋ does

TABLE 1 | A summary of the languages with the ISO codes, language families, the number of speakers [39], and regions in East Africa where the languages are mainly spoken.

Language	ISO 639-2 code	Family	Speakers	Region
Luganda	lug	Niger-Congo-Bantu	7M	Central and Southern Uganda
Runyankore-Rukiga	nyn	Niger-Congo-Bantu	3.4M	Western Uganda
Lumasaba	lus	Niger-Congo-Bantu	1.6M	Eastern Uganda
Acholi	ach	Western Nilotic	1.5M	Northern Uganda and Southern Sudan
Swahili	swa	Niger-Congo-Bantu	98M	Central and East Africa

not appear on computer keyboards and is often replaced by the combination ng'.

3.2 | Runyankole-Rukiga

(nyn) are two languages that belong to the JE10zone of the Great Lakes, Narrow Bantu of the Niger-Congo language family. Runyankole-Rukiga are mildly tonal and highly agglutinating with a large noun class system [41]. They exhibit high incidences of phonological conditioning that make them complex to deal with computationally.

3.3 | Lumasaba

(lus) is a Bantu language under the Niger-Congo family and is mostly spoken in the Masaba or Bugisu region in Uganda. Lumasaba has two main categories of dialects: the northern forms such as Ludadiri, Luhugu, and Luwalasi and the southern forms such as Lubuya, Lusoba, and Lukiende. Text in all Lumasaba dialects is represented using the Latin alphabet without two letters (Xx) and (Qq), but also with (used to represent voiced velar nasal which is common in most Bantu languages) [42].

3.4 | Acholi

(ach) belongs to the Lwoo languages under the Nilotic family of languages [43]. Acholi is mostly spoken in Northern Uganda in the districts of Gulu, Kitgum, Nwoya, Lamwo, Amuru, Pader, and some parts of Southern Sudan. Text in Acholi is represented using the Latin alphabet with the exception of seven letters (Ff, Hh, Qq, Ss, Vv, Xx, and Zz), but also with character.

3.5 | Swahili

(swa) also known as Kiswahili is the most widely spoken language on the African continent. Swahili belongs to the Bantu family of languages and it distinguishes itself from other Bantu languages by its vocabulary of both Bantu and Arabic origins. It has 5 vowels (a, e, i, o, u), 19 consonants (b, c, d, f, g, h, j, k, l, m, n, p, r, s, t, v, w, y, z [exclude x and q]) and 9 digraphs (ch, dh, gh, kh, ng', ny, sh, th, ng) unique to Swahili pronunciation [44].

4 | Text Data Creation

Our data collection process followed a participatory approach which was provided based on previous research in [4, 32]. The process also involved several stakeholders in the data collection, curation, model building, and evaluation processes. We had native language speakers across the East African countries as *content creators*. The content creators helped us to produce language-specific text data. The content was available in both digital and nondigital forms [6]. The *translators* provided the relevant translations across the four East African languages. The translators were diverse from the linguists, researchers, and professional translators. The project also worked with *curators* who were individuals involved in content selection for the specific East African languages as we will outline in this section. Finally, we also worked with a diverse group of *annotators* who enabled us to build out the sentiment-tagged dataset in two languages.

4.1 | Data Sources

We started with building out the English corpus which was later translated to four Ugandan languages and one East African langauge. The English text corpus were created from different sources to enable us to capture diversity and local context. The content creators used a number of text sources for the base English corpus, for example, English Wikipedia, social media, online local newspapers, storybooks, novels, human rights charters, and so forth. To create the monolingual corpus, data were sourced from the community through a crowdsourcing approach and the collection of preexisting text from several sources such as Wikipedia, storybooks, novels, Luganda teacher guides, radio transcripts, Luganda and Swahili Bible translations and reports, by the curators. The sentences collected through a crowdsourcing approach were obtained via a web-based application where they were validated and reviewed by the curators. We also engaged linguists who created sentences based on different topics of interest. The community contributions included language text made by linguists and native language speakers. Owing to the limited access to technology, the content creators from the community were not able to provide digital content [4]; instead they wrote their contribution on paper, which we had to type out.

4.2 | Data Preprocessing and Preparation

Given the variety of data sources from which the text data were sourced, that is, structured (mass media, online) and unstructured sources (social media), it was essential to preprocess the data. This activity involved data cleaning and sentence generation activities. The social media text was cleaned using a Python script to remove any unwanted characters such as URLs, email addresses, mentions, and hashtags. As previously discussed, we had Luganda sentences generated by the content creators, which were partly handwritten. As part of the preprocessing step, the sentences were typed out and eventually were reviewed and validated.

The English sentences that were sourced were used as prompts to generate new sentences. We preferred not to use these original sentences directly for translation across multiple languages as there can be copyright issues and also we had to simplify the formal tone of the text [31]. This task was also important because the data would eventually be used to create the speech corpus that required that the sentences be readable. We developed a set of guidelines that were followed during the creation and review of the new English sentences:

- Sentences had to retain the context provided by the original sentence.
- Sentences had to contain 4–14 words.
- Sentences had to be syntactically and semantically correct.
- Words in the sentences need to be simple and easily pronounceable.
- Sentences should not contain abbreviations and numbers.

TABLE 2 | Overview of the dataset creation process. The sentence on the left-hand side is the original sentence while the sentence on the right-hand is the created sentence based on the sentence prompt. The translations across the five languages are also shown.

Original sentence	New sentence	Language	Translations
Ebola outbreaks are common in DRC.	There is an outbreak of Ebola virus disease in Uganda.	Acholi	Two Ebola opoto i Uganda.
		Luganda	Ekirwadde ky'akawuka ka Ebola kibaluseewo mu Uganda.
		Lumasaba	Luffu lwa eboola lwakwilewo mwinambo.
		Runyankole-Rukiga	Hariho okubaruka kw'oburweire bwa ebora omuri Uganda.
		Swahili	Kuna milipuko ya ugonjwa wa virusi vya Ebola nchini Uganda.

Table 2 provides an example of a sample sentence created based on the original sentence and the subsequent translation of the new sentence to Acholi, Luganda, Lumasaba, Runyankole-Rukiga, and Swahili.

4.3 | Sentence Creation

After the text data creation step, the next step was to translate the sentences from English to the five East African languages, a task that was carried out by a team of language-specific translators. The translation carried out for the Ugandan languages was completed using both online and offline tools, whereas the Swahili translation is done offline. As an online translation tool, Pontoon,[5] a translation management system developed by Mozilla was used to facilitate the text translation process. The translation across all languages was divided into two phases: (1) contribution, in which contributors provided translations and (2) validation, in which language validators reviewed and validated the contributors' translations.

We created text datasets that include 40,000 English text corpus that was translated into five East African languages: Acholi, Runyankore, Luganda, Lumasaba, and Swahili, to create a parallel corpus. Table 3 provides examples of the dataset with translations from English to the five East African languages. Table 4 provides a summary of the number of sentences in each of the translated language pairs.

4.3.1 | Monolingual Corpus

We created a monolingual corpus for the five East African languages. For Lumasaba, text data were collected in both dialect categories. Table 5 provides an overview of the size of the monolingual corpus created for each of the five languages.

4.3.2 | Sentiment-Tagged Dataset

We created a sentiment-tagged dataset that was translated into Luganda and Swahili. Table 6 provides examples of the dataset with translations from source English sentences to Luganda and Swahili and the annotated sentiment. The dataset defines two annotation classes: positive (POS) sentiment and negative (NEG) sentiment. A sentence was tagged with a positive sentiment if it implied a positive attitude, emotion, or sentiment whereas a sentence was tagged with a negative sentiment if it implied a negative attitude, emotion, or sentiment. The sentiment-tagged parallel corpus for Luganda and Swahili is available in [45].

4.4 | Guidelines for Monolingual Corpus Creation

We created a monolingual dataset for the five East African languages (Acholi, Luganda, Lumasaba, Runyankore-Rukiga, and Swahili). The task involved the creation of new sentences and curating sentences from various sources into an acceptable format. The sentence creation and review followed several guidelines summarized here. The sentences had to be of a specific length, preferably between 4 and 30 words. The sentences did not contain abbreviations to avoid ambiguity. The sentences had to be syntactically and semantically correct. The sentences had to be written using the proper orthography, and all digits had to be written in words. The symbols and non-alphanumeric characters had to be excluded apart from full stops and commas. It was a requirement that sentences had to contain only one language except for a few terms or nouns adopted from other languages. Although many speakers in East Africa code switch as they speak, the focus is only on one language per sentence. The sentences had to contain or have a local context, for example, a country context with local entities such as names of people, places, dates, and organizations, and a part of the dataset was used for building a sentiment-tagged dataset. This task involved the assignment of either a negative or a positive sentiment to the sentences. The sentences created had to be meaningful and with connotations that can be associated with a positive, negative, or neutral label.

4.5 | Guidelines for Speech Corpus Creation

Part of the monolingual corpus for Luganda and Swahili was used to collect voice contributions on the CV platform. The text

TABLE 3 | Sample examples of the parallel translated sentences.

English	Acholi	Luganda	Lumasaba	Runyankole-Rukiga	Swahili
Government should fully equip the labor wards in most public hospitals.	Gamente myero oket jami mamite ducu i ot nywal i pol odi yen pa gamente.	Gavumenti esaanye eteeke ebikozesebwa mu mazaaliro agasangibwa mu malwaliro ga gavumenti.	Linambo lyakha khura ibyo byosi bibikanibwa mumakangilo nadala isi bamayi basalila mumakangilo ke babandu.	Gavumenti eshemereire kuta ebikozeso byona omu marwariro g'abazaarisa aga gavumenti.	Serikali inapaswa kuandaa kikamilifu wodi za kazi katika hospitali nyingi za umma.
Family businesses hardly survive beyond the founders.	Pol biacara me gang pe bedo ki inge lucakkone.	Bbiizinensi ez'ab'enganda tezitera kubeerawo nga abaazitandikawo bamaze okufa.	Kamakulano ke mungo sikatala khutuuma baba kananikhawo taa.	Bizinesi z'ab'ekika tizirikukira kugumizamu bwanyima y'okufa kw'abaazitandikire.	Biashara za familia haziishi zaidi ya waanzilishi.
Every person above eighteen years has a right to vote.	Ngat mo keken ma mwakane kato apar wiye aboro tye ki twero me bolo kwir.	Buli muntu asussa emyaka ekkumi n'omunaana alina eddembe okulonda.	Bbuli mundu ulli ni kimiko likhumi nasinane abba ni ikifunisi ye khurobola.	Buri muntu orikurenzya aha myaka ikumi na munaana aine obugabe bw'okuteera akaruuru.	Kila mtu aliye na umri wa zaidi ya miaka kumi na minane ana haki ya kupiga kura.
The proposal focuses on reducing the tax on mobile money withdrawals.	Tam ma kikato kwedeni neno kit me dwoko piny wel mucoro me kwanyo cente ki i mobile money.	Ekiteeso kiruubirira kukendeeza musolo ogw'okuggyayo ensimbi ng'okozesa mobile money.	Sitesoo silolelela khukhaala kumusolo kuli mu khurusayo kamapesi khu siimu.	Enteekateeka egyendereire kutuubya aha mushoro gw'okwiha sente aha simu.	Pendekezo hilo linalenga kupunguza ushuru kwa pesa za rununu.
The government has not done enough to stop illegal land grabbing.	Gamente pe otiyo ma oromo me juku tim me kwanyo ngom i wi alii.	Gavumenti tekoze kimala okumalawo obubbi bw'ettaka obumenya amateeka.	Linambo silya kholile sikale khukhwakamisa khuyutula kamaswa khukhali mumakambila taa.	Gavumenti tekozire kihango kwihaho obwibi bw'amataka.	Serikali haijafanya vya kutosha kuzuia kunyakua ardhi haramu.

dataset was further reviewed to ensure that the sentences were speakable. The sentences had to be broken into two sentences if they were longer than 14 words. It is essential that abbreviations and acronyms were avoided and there were no digits in the sentences. The words used needed to be simple and easy to understand. It was also important that the sentences had no form of religion, culture, or ethnicity biases.

4.6 | Guidelines for Sentence Translation

In this task, we translated sentences from English to the five East African languages to create a parallel corpus. The task was carried out by a team of language experts using these guidelines. The translators had to ensure that the context of the sentences was preserved during translation. The experts had to ensure that they understood the meaning of each English word before translation. If the source sentence contained digits or acronyms, these had to be preserved in the translation. Finally, if the numbers were written out in words, these also had to be written out in words in the translation.

TABLE 4 | Parallel text corpus.

Parallel corpus	
Language	**Dataset size**
English—Luganda	40,000
English—Acholi	40,000
English—Lumasaba	40,000
English—Runyankore	40,000
English—Swahili	40,000

TABLE 5 | Monolingual text corpus.

Monolingual corpus	
Language	**Dataset size**
Luganda	400,000
Acholi	40,047
Runyankore	40,211
Lumasaba	40,184
Swahili	200,000

5 | Speech Dataset Collection

We created speech data for two commonly used languages in East Africa, that is, Luganda and Swahili. The speech dataset was collected using Mozilla's CV crowdsourcing platform [46]. The CV platform is a project aimed at collecting free, public language datasets while guaranteeing that language data accurately reflect the diversity of people. People can donate their voices on the CV platform by reading and recording sentences, and they can also listen to and validate the accuracy of other people's recorded audio clips.

5.1 | Community Crowdsourcing

The Luganda voice data contribution in Uganda was carried out through a community crowdsourcing approach from communities made up of university student groups and the Association of Luganda Language Teachers. The community groups would gather physically to record voice clips and validate the recorded voice clips on the CV platform. However, due to the COVID-19 pandemic and lockdown measures imposed, voice contributions and validations were conducted online. The contributors were given weekly internet bundles to compensate for voice contribution and validation. At the end of each month, we also provided cash rewards to the top three voice contributors, voice validators, and community mobilizers to further encourage voice contribution and validation.

After the CV mobilization, it was observed that the Luganda voice dataset was not very diverse regarding age, dialects, and accents. The mobilization strategy was then revised to include more contributions from the community with an age range of above 40 years. This mobilization was carried out in the different counties of Buganda Kingdom. The voice data collection drives aimed to collect voice contributions from people with various dialects, accents spoken in various counties, and age diversity. This also enabled us to achieve the aim of preserving and documenting the Luganda's linguistic diversity. The county chief led the mobilization and recruitment of the community within each country.

The Swahili voice contribution in Kenya and Tanzania was also done through a community crowdsourcing approach across universities in these countries. Students from various universities, such as Maseno University and Kabarak University in Kenya,

TABLE 6 | Sample examples of the translated sentences and annotated sentiments.

Source sentence	Luganda translation	Swahili translation	Sentiment
My pastor loves visiting the sick in the hospital over the weekend.	Omusumba wange ayagala okukyalira abalwadde mu ddwaliro ku nkomerero ya wiiki.	Mchungaji wangu anapenda kwenda kuwaona wagonjwa hospitalini mwishoni mwa wiki.	Positive
We were disappointed when you refused to reply to our messages.	Kyatwewuunyisa ng'ogaanye okuddamu obubaka bwaffe.	Tulikata tamaa ulipokataa kujibu ujumbe wetu.	Negative
Thanks to the government of Uganda's fight against the spread of coronavirus.	Gavumenti ya yeebale kulwanyisa nsaasaana y'akawuka ka kolona.	Pongezi kwa serikali ya Uganda kwa mapambano dhizi ya kusambaa kwa virusi vya Uviko.	Positive
People yelled and cried as the thieves burnt their goods to ashes.	Abantu baajogoolikana era ne bakaaba ng'ababbi bookyezza ebintu byabwe ne bisiriira.	Watu walipiga walilia na kupiga kelele walipochomewa vitu vyao na wezi mpaka majivu.	Negative

the University of Dodoma, Nelson Mandela African Institution of Science and Technology, and Dar es Salaam School of Journalism in Tanzania, participated in the voice contribution and validation processes.

5.2 | Voice Contributions

The speech data generated for the CV project was measured by the number of recorded and validated hours. Since the launch of Luganda on the CV platform in September 2020, Luganda has recorded a total of 582 hours and 438 validated hours from over 646 unique contributors based on Common Voice Corpus 15.0.[6] The voice contributors were people from rural to urban settings, educated and noneducated, and low- and middle-income earners. We had 35.80% male and 38.73% female voice contributions, 230 male voices and 222 female voices, giving us a female and male ratio of voices of 0.965. For the Swahili language, a total of 1100 hours were recorded, 393 hours validated from 1103 unique voices, and 86,484 sentences were created based on Common Voice Corpus 15.0.[7] We had 36.07% male and 32.37% female voice contributions, 278 male voices and 262 female voices, giving us a female and male ratio of voices of 0.942. Our voice contributors' age ranged between 18 and 80 years. Table 7 summarizes the age demographics of the contributors for Luganda and Swahili.

6 | Software Tools

This section describes the software tools used for MT, sentence crowdsourcing, and voice crowdsourcing tasks.

6.1 | Pontoon Localization Platform

Sentence translation from English to Luganda, Acholi, and Runyankole was carried out using the Pontoon platform, a translation management system developed by Mozilla.[8] To facilitate the translation process, new locale instances were created and languages were added, each to a specific instance. English sentences were uploaded to the system, and users were created and allocated to different locales. The users could view English sentences and translate the sentences to a language as specified in the locale instance. The system also facilitated the review of the contributed sentences. With this feature, a user could read both sentences (the English sentence and the translated version of the sentence) and then either approve or reject the translation. A dataset for the approved sentences was later downloaded and used to train MT models. Figure 1 shows an example of translation results for three Ugandan languages from the Pontoon system.

TABLE 7 | Demographics of common voice corpus for Luganda and Swahili languages.

	Language	
Age group	**Luganda (%)**	**Swahili (%)**
18–19	0.7	0.1
20–29	37.7	44.3
30–39	23	12.3
40–49	7.2	3.9
50–59	4.5	4.7
60–69	1.7	0.7
70–79	0.2	—
80–89	0.03	—
90–99	0.32	—
Unknown	25.6	33.95

6.2 | Sentence Crowdsourcing Tool

A web application was developed to collect Luganda monolingual sentences using the crowdsourcing approach. With this approach, a group of linguists was sourced to create the Luganda sentences. The application features include the following: (a) *sentence categorization* feature where contributors were allowed to create sentences under different topics of interest, for example, education, sports, politics, entertainment, COVID-19, and agriculture; (b) *sentence validation* feature where the application facilitated sentence validation to ensure the created or uploaded sentences met the guidelines defined in the sentence creation guidelines. Each new instance was compared with all the existing sentences in the database to ensure no duplicates, and the (c) *bulk upload* feature where the application allowed bulk sentence upload. The sentences were validated using the defined rules in the sentence creation guidelines.

For the Swahili sentence contribution, a web-based tool was created to enable users to submit original sentence creations in either text or CSV format. This ensured that users could participate actively at various times across different places. Figure 2 shows an interface where the user could either select a topic of interest, type a sentence, then submit the sentence or upload a word document containing the sentences organized in a paragraph manner.

6.3 | Mozilla's CV Platform

After validating and reviewing the sentences collected from the crowdsourcing tool shown in Figure 2, the sentences were uploaded to the CV platform to enable voice contribution and validation. Figure 3 shows an interface where the user can either decide to donate their voice or validate the available voice clips.

6.4 | Yogera Application

We built a new mobile and web application for voice data collection and validation. The reason for developing this platform was the restrictions on CV on the nature of sentences uploaded. The CV platform only allows sentences under the CC-0 Public Domain license. However, for sentences that could not meet the license requirements, we developed a mobile application that provided a flexible licensing structure for content creators. There was also a need to make voice contribution and validation easier for the contributors via a mobile application. Figure 4 shows an interface where users can either donate their voices or validate the contributed voice clips.

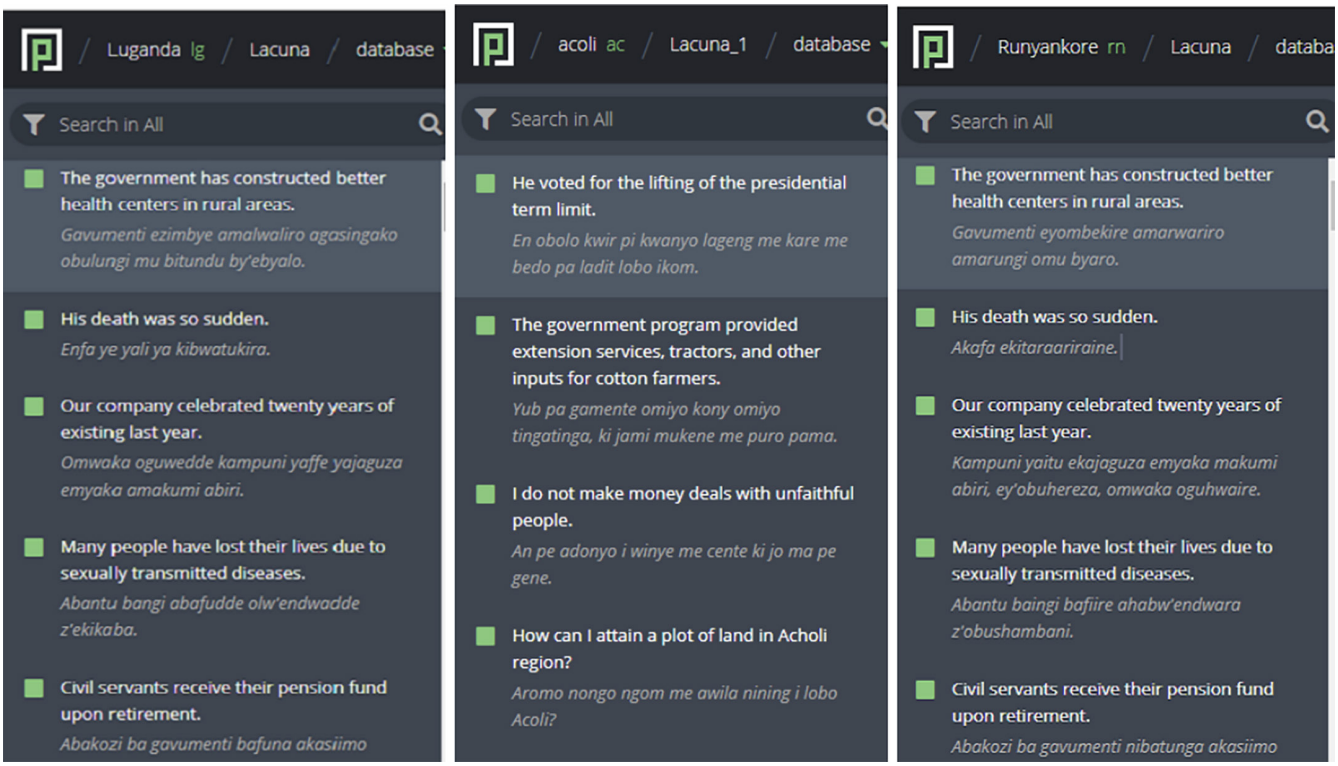

FIGURE 1 | Screenshot showing interfaces of the Pontoon tool with the sentence translations from English to Luganda, Acholi, and Runyankole.

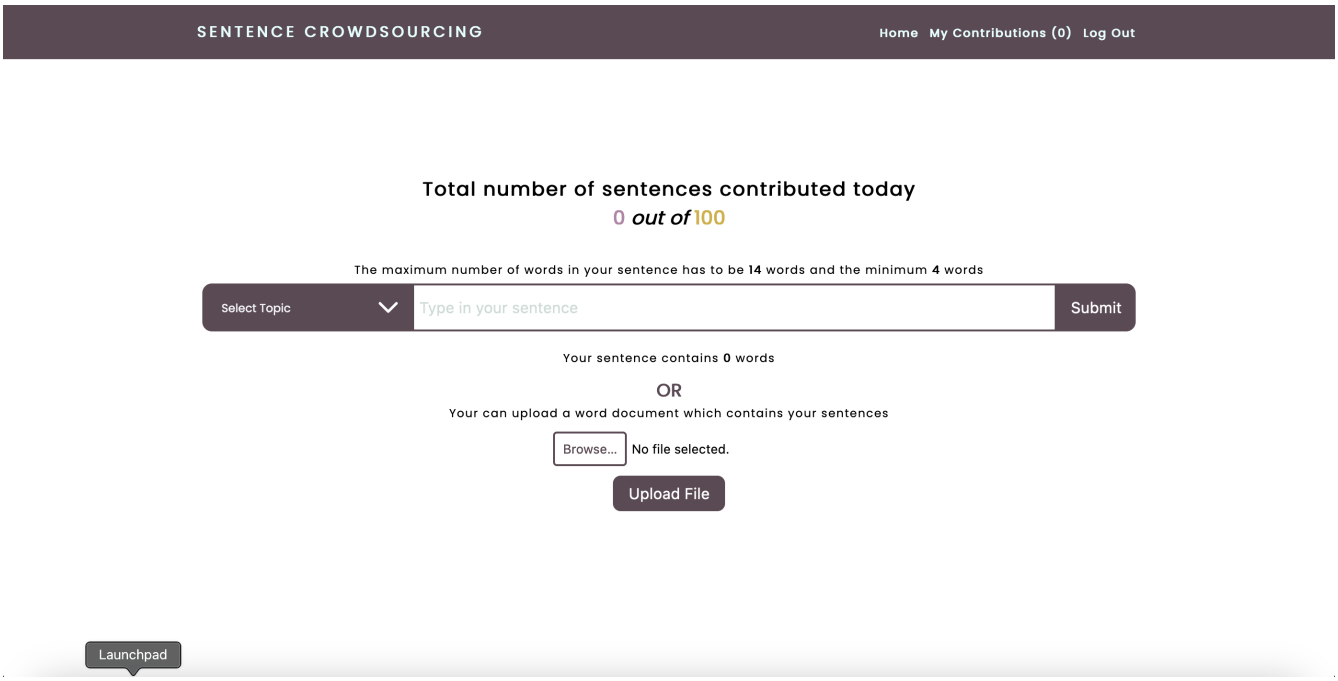

FIGURE 2 | Screenshot showing an interface of the sentence crowdsourcing tool.

7 | Quality Assurance

A linguist expert reviewed and validated each translation to ensure a high-quality dataset was obtained. It was observed that some English words can have multiple meanings in local languages, which could lead to multiple translations. The linguists handled this by choosing the most appropriate translation based on the context of the sentences. Linguists from the language-speaking communities led by the language coordinator provided quality control through discussion and fixtures of problematic translations. They also performed multiple checks to find and correct the misspellings in the dataset, which is a similar approach in other translation projects [26]. Each language had a lead responsible for ensuring the quality and approval of the final translations.

To further evaluate the quality of the parallel text corpus, we ran baseline models for MT. We evaluated the performance of

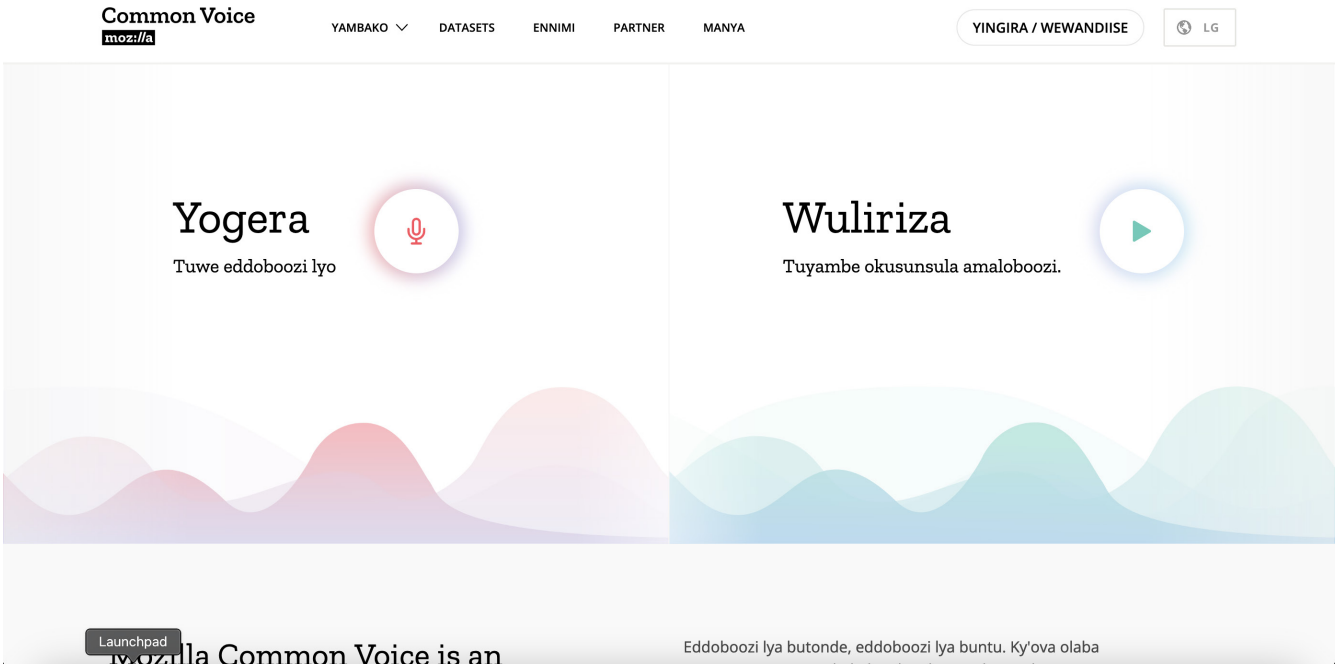

FIGURE 3 | Screenshot showing an interface for the Mozilla's common voice platform.

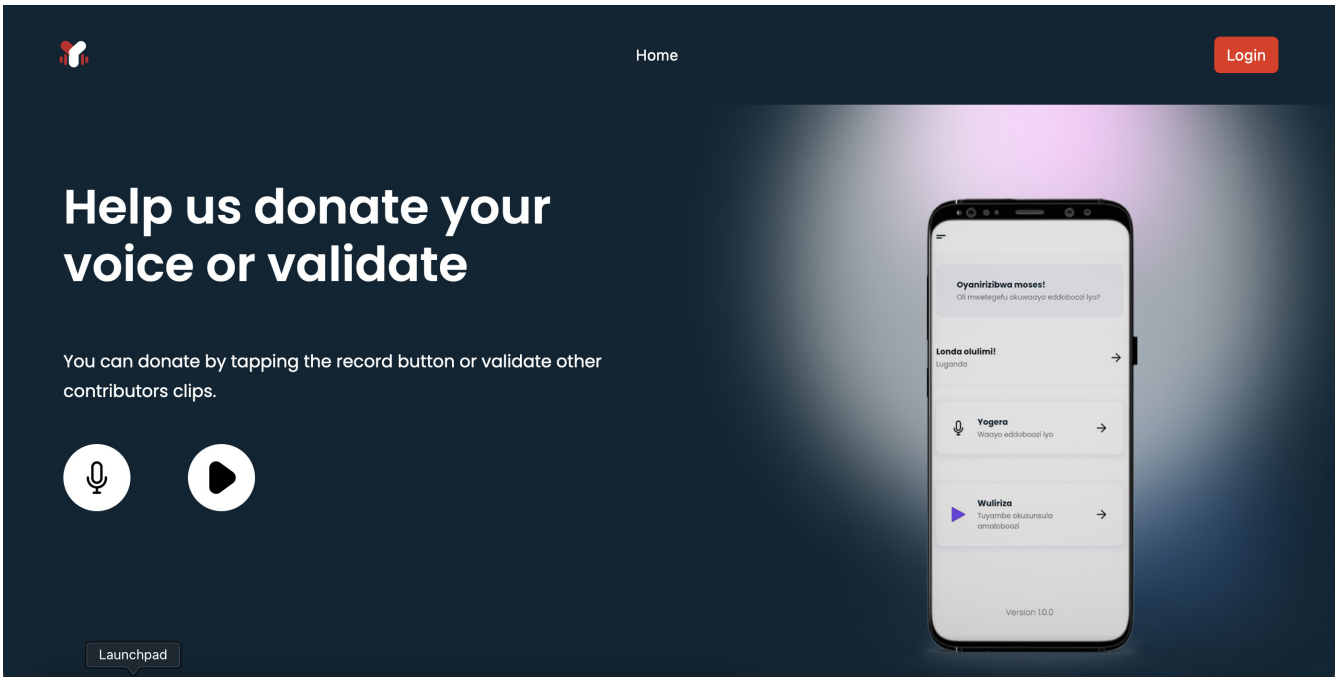

FIGURE 4 | Screenshot showing an interface for the Yogera web platform.

the models using BLEU score, a metric widely used in comparing machine-translated text to a set of high-quality translations. The results from these experiments are presented in Section 8.1. The performance of our model is comparable with the model developed on the SALT dataset [31]. During the creation of the sentiment-tagged dataset, two people reviewed and validated each data point. At the last layer, the human sentiment tag was compared with a machine-generated sentiment tag, which had been returned with a confident score between 0 and 100. On average, the confidence score across the sentiment dataset was 95%.

8 | Baseline Models

This section discusses results from the baseline models based on the text and speech datasets. The downstream tasks included MT, topic modeling and classification, sentiment analysis, and ASR.

8.1 | Machine Translation

The experiments for the English-to-Luganda machine are based on the English-to-Luganda parallel corpus. The dataset was split into train (80%), test (10%), and validate (10%) sets. We used

63,840 parallel sentences in the train, 10,640 in the test, and 10,640 in the validation sets, respectively. We created translation objects that write JavaScript object notation (JSON) files of the train, test, and validation sets. The JSON files were pushed to the Hugging Face platform,[9] providing a collaborative platform for model training. We trained a transformer model [47] specifically, the multilingual Marian MT model for the English–Luganda translation tasks [48]. We leveraged transfer learning on the *Helsinki-NLP/opus-mt-lg-en* and *Helsinki-NLP/opus-mt-en-lg* pretrained multilingual models. The models were trained for 30 epochs with a batch size of 16 and 10,640 sentences from the validation set at each training step. Our baseline results demonstrated good performance on the test set and a translation quality with a BLEU score of 26.0 for the English–Luganda model and a BLEU score of 24.6 for the Luganda–English model as shown in Table 8.

8.2 | Topic Modeling and Classification

This section describes the topic modeling and classification models and results based on the Luganda monolingual corpus. The dataset includes 20,000 sentences categorized across 14 major topics, as shown in Table 9. We also had sentences that did not belong to any mentioned topic, and these were labeled as "Others."

8.2.1 | Topic Modeling

Under the topic modeling, we used the nonnegative matrix factorization (NMF) technique to discover meaningful topics from the unlabeled monolingual Luganda corpus. NMF model considers each sentence as a data point which is taken as a high-dimensional vector. The NMF model decomposes a high-dimensional vector into low-dimensional representations that are nonnegative matrices. Using the original matrix (A), NMF gives two nonnegative matrices where W the basis matrix represents the topics the model found and H is the coefficient (weight) of a topic [49]. Equivalent to this is that a matrix holds records by words, an H matrix holds records by topics, and W is a representation of topics by words. The NMF model was trained using features extracted from a TF-IDF vectorizer. The TF-IDF vectors are high-dimensional representations of the input data that allow the NMF model to adjust the initial values of the factor matrices W and H to minimize the reconstruction error between the input data and its factorization. The hyperparameters of the NMF model were set to their default values, except for the solver, which was set to "mu," the maximum number of iterations set to 1000, the regularization parameter "alpha" set to 0.01, and the l1-ratio set to 0.5.

Table 10 shows the meaningful topics and the top 10 words that are associated with each of these topics as generated from the NMF model. The model clustered some of the topics in more than one cluster which included Land ("Ebyettaka") in topics 5 and 10, Security ("Ebyokwerinda") in topics 6 and 12, Politics ("Ebyobufuzi") in topics 14 and 15, Education ("Ebyenjigiriza") in topics 8 and 21. These topics carried more weight than the other topics in the dataset.

TABLE 8 | Machine translation evaluation metrics on the test set.

Translation direction	BLEU
eng → lug	26.0
lug → eng	24.6

TABLE 9 | Topic names in English and Luganda and their respective topic codes.

Topic name	Topic code
COVID-19 ("Kolona")	Covid
Security ("Ebyokwerinda")	SE
Agriculture ("Ebyobulimi")	Agri
Culture ("Ebyobuwangwa")	C
Transport ("Ebyentambula")	T
Environment ("Ebyobutonde")	Env
Politics ("Ebyobufuzi")	P
Health ("Ebyobulamu")	H
Religion ("Eddiini")	R
Sports ("Ebyemizannyo")	S
Business ("Ebyenfuna")	B
Land ("Ebyettaka")	Land
Legal ("Amateeka")	L
Education ("Ebyenjigiriza")	Educ

8.2.2 | Topic Classification

During data exploration, we observed that the data were unevenly distributed across the different topics, which could lead to bias among standard classifiers [50]. Using random sampling, we created a balanced dataset across all the topics. According to random sampling, samples of the data were randomly taken from the minority classes and duplicate instances were created so that the minority class reaches a size comparable with the majority class [50]. We developed Luganda word embeddings, that is, FastText, Glove, and paragram, and word2Vec embeddings from the monolingual Luganda dataset using Gensium[10] and Glove frameworks.[11] The developed Luganda word embeddings were used in the feature processing to perform model training for Luganda topic classification models.

We trained baseline, neural network, and pretrained models for the topic classification task.

- **Baseline models**: These were trained using a count vectorizer to tokenize the input data. This method builds a vocabulary of known words and encodes new documents using that vocabulary. The baseline models that were implemented included Logistic Regression, Support Vector Machines, Multi-Layer Perceptron, and XGBoost using the scikit-learn framework.[12] The dataset was split into a training set and a testing set with a ratio of 9:1.
- **Neural network models**: We input tokenized sentences that were converted into vectors, padded, and truncated to ensure a fixed length. The resulting padded sequences for each sentence were used as input to the model to predict the corresponding topic class. The dataset was split into

TABLE 10 | Latent topics and the top 10 words derived from the NMF model.

Topic 2	Topic 5	Topic 6	Topic 8	Topic 9	Topic 10
uganda	ettaka	poliisi	abaana	abayizi	ettaka
bukiikakkono	enkaayana	emisango	ssomero	ssomero	ensonga
amawanga	ensonga	omusango	balina	abasomesa	lyabwe
ekitongole	ssentebe	abateeberezebwa	baali	akamalirizo	okubumbulukuka
nsi	lyabwe	okunoonyereza	abazadde	bubi	enkaayana
mulimu	zirina	okwekalakaasa	abato	masomero	kitundu
afirika	ekyalo	kaduukulu	okugenda	ensoma	ttaka
enjawulo	lisobola	ekitundu	essomero	bajja	abalimi
ababudami	okuwandiisa	okukuuma	abawala	omusomesa	nsonga
kenya	mingi	ekwata	ewaka	abali	eryo
Topic 12	**Topic 14**	**Topic 15**	**Topic 16**	**Topic 17**	**Topic 18**
poliisi	pulezidenti	okulonda	kitundu	enjawulo	enguudo
amateeka	obwa	omwaka	obutebenkevu	obuwangwa	embi
omusango	kkampeyini	ogujja	obutali	ebifo	obubenje
emisango	okulonda	pulezidenti	mugaso	amawanga	zirina
bantu	kwa	bwenkanya	emirembe	bulina	nguudo
kitundu	ajja	ennaku	obuwangwa	enzikiriza	ennungi
ekitongole	myaka	omuntu	amagye	byobuwangwa	nnyingi
abamenyi	emyaka	kwaliwo	butebenkevu	ngeri	mbeera
kkooti	omwaka	okuba	entambula	ebintu	ensimbi
eggulo	ekkomo	kujja	ebyokwerinda	bingi	entambula
Topic 19	**Topic 20**	**Topic 21**	**Topic 22**	**Topic 24**	**Topic 25**
emiti	katonda	abaana	ekibiina	bangi	ssente
obutonde	kkanisa	okusoma	omukulembeze	ekirwadde	okukola
ensi	omuntu	abazadde	ebyobufuzi	ssenyiga	bizinensi
okutema	obulamu	ssomero	ekibinja	omukambwe	okufuna
okusimba	alina	masomero	abakulembeze	abavubuka	pulojekiti
okukuuma	abakrisito	bateekeddwa	abayeekera	obulamu	nnyingi
ensigo	okubeera	beetaaga	eky	bafudde	oluguudo
butebenkevu	ngeri	amasomero	oluvuganya	ggwanga	emirimu
kyonoona	kisa	engeri	kino	bwabwe	ekitundu
batema	okuweereza	abato	kitundu	ababbi	kkampuni

train, validation, and test sets with a ratio of 8:1:1. We experimented with bidirectional LSTM with 2D Maxpooling proposed by Peng et al. [51], with default hyperparameters. However, we hyper-tuned the original feature map to 250 and a min-batch to 64 because we had a small dataset. We also experimented with gated recurrent unit (GRU) proposed by Junyoung et al. [52] with default hyperparameters. For both the bidirectional and GRU models, word embeddings were applied to the first layers.

- **Pretrained models**: We trained the BERT and RoberTa models with the transformer model architecture pretrained on English data for text classification from the Hugging Face platform [47]. We trained the models using the Luganda dataset and the dataset was split into train, validation, and test sets with a ratio of 8:1:1.

The performance of each of the models was evaluated using the F1 score, precision, and recall metrics. These metrics were computed based on the values present in the confusion matrix, which included true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs), as illustrated in Figures 5 and 6.

Table 11 shows the results for all classifiers used in this analysis. The best performers are SVM, a classical approach, and BERT, a pretrained model.

8.3 | Sentiment Classification

The sentiment classification experiments were conducted for the Luganda sentiment-tagged dataset. The dataset

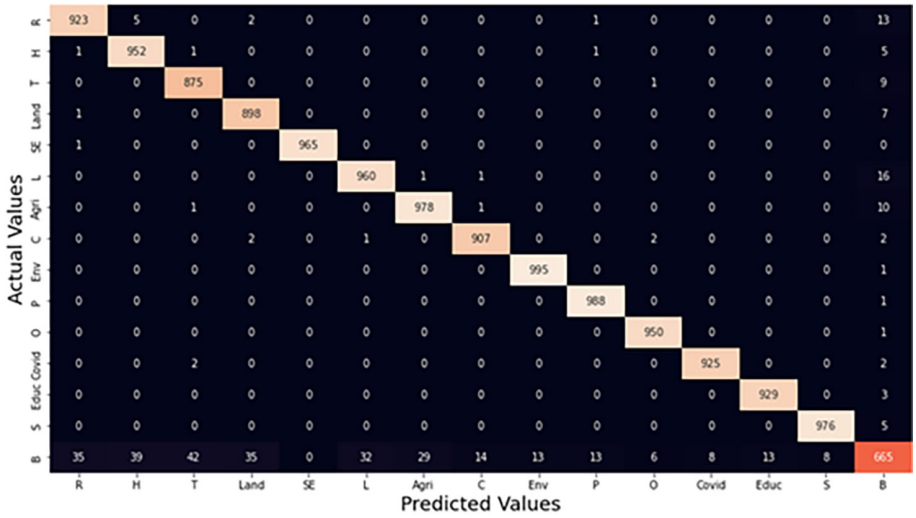

FIGURE 5 | Results for the BERT classifier.

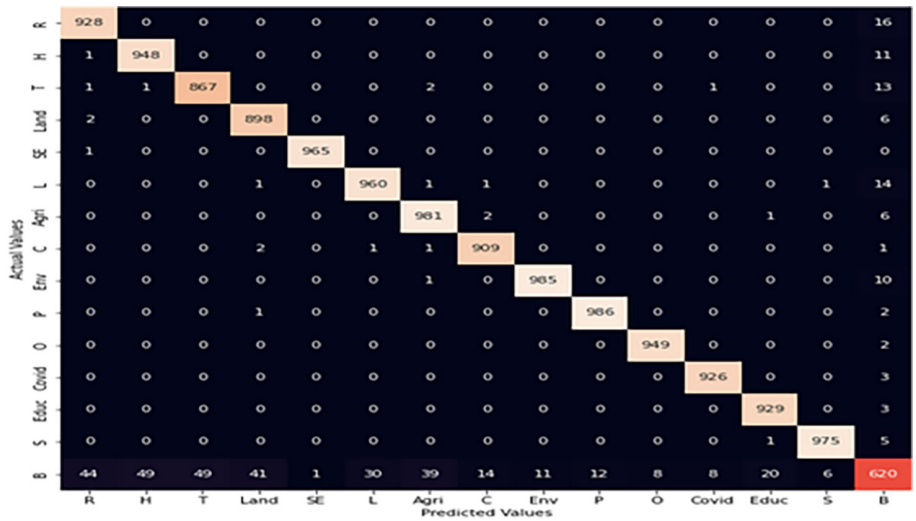

FIGURE 6 | Results for the SVM classifier.

TABLE 11 | Results for topic classification models.

Classifier	Precision	Recall	F1 score
XGB classifier	0.873	0.852	0.859
Logistic regression	0.947	0.949	0.948
SVM	0.967	0.969	0.967
MLP classifier	0.972	0.971	0.973
Bi LSTM	0.826	0.822	0.823
GRU	0.833	0.829	0.830
RoberTa	0.974	0.975	0.975
BERT	**0.983**	0.983	0.983

Note: Bold show the results of the best performing models.

included the positive and negative classes which were evenly distributed. The data were trained on the baseline models including the multinomial Naive Bayes, stacking classifier, and the neural network models. We used a count vectorizer to convert the data into tokens to be used by the baseline models. The data were divided into train, validation, and test splits in an 8:1:1 ratio. We also trained a bidirectional LSTM with default hyperparameters [51] to perform sentiment analysis on the Luganda dataset. The performance of each of the models was evaluated using recall, precision, and F1 score. Table 12 shows the performance of the different classifiers on the sentiment classification task.

8.4 | Automatic Speech Recognition

To build the STT models, we used 300 hours of validated voices collected using the CV platform. Each entry in the CV dataset consists of a unique MP3 and corresponding text file. Part of the recorded hours in the dataset also included demographic metadata such as age and sex. The dataset comes with a clips folder, invalidated.tsv, reported.tsv, train.tsv, dev.tsv, other.tsv, validated.tsv, and test.tsv splits. The splits are done using the Mozilla Corpora Creator.[13] For our experiments, we used the Coqui STT toolkit[14] to build the STT models. Coqui STT expects audio files to be WAV format, mono-channel, and with a 16 kHz sampling rate and yet the dataset contains MP3 audio.

TABLE 12 | Model performance for the sentiment analysis models.

Classifier	Precision	Recall	F1 score
Stacking classifier	**0.90**	0.90	0.90
Multinomial NB	0.88	0.88	0.88
LSTM	0.84	0.83	0.83

Note: Bold show the results of the best performing models.

TABLE 13 | Word and character error rates on the test set.

Model	WER	CER
Coqui STT	23	9
XLSR-wav2vec2	**12**	**3**

Note: Bold show the results of the best performing models.

Using the CV importer that comes with Coqui STT, we preprocessed all the CV data to .csv files (train, dev, and test), and all the audio files were converted to .wav format. The .csv file output had a format of *wav_filename*, *wav_filesize*, and *transcript*. The *wav_filesize* in bytes is used to group audio of similar lengths for efficient batching. We used the English alphabet as our output alphabet. Using the commonvoice-utils[15] package, we did basic linguistic checks to identify characters not defined in the alphabet. Our approach utilized Coqui STT model based on Baidu's Deep Speech architecture [53]. We used a cross-lingual transfer learning approach by initializing the parameters of Coqui STT's neural network from a pretrained English model and then fine-tuning these parameters to Luganda. The English release model was downloaded and fine-tuned on Luganda data. We trained the Luganda model for 100 epochs with a dropout of 0.2, a batch size of 48, and a learning rate of 0.001.

We used a probabilistic language model to build a scorer for our acoustic model. Using a KenLM toolkit [54], we built a 5-gram language model. We use a text corpus of 240,000 sentences for our first language model and initial tests. The sentences used were from the Luganda monolingual corpus created in this project. The corpus was cleaned with one sentence per line. We evaluated the model on a CV test corpus of 20.3 hours using a model scorer to obtain a word error rate (WER) of 23%. These results are built on the previous work we had done in building STT models for COVID-19 monitoring [36].

The second model we utilized was Facebook's XLS-R wav2vec2, a large-scale model for cross-lingual speech representation learning based on wav2vec 2.0 [55]. It is pretrained on 436k hours of unlabeled speech, including VoxPopuli, MLS, CommonVoice, BABEL, and VoxLingua107. The model uses the wav2vec 2.0 objective in 128 languages. There are three pretrained XLS-R models, each differentiated by the number of parameters. We used the Hugging Face pipeline to load the 300M pretrained model[16] and fine-tuned it on Luganda data. We trained the model for 50 epochs with a learning rate of 0.0003, per device batch size of 16, and eight gradient accumulation steps. We boost the model using the 5-gram language model, which was discussed earlier. The model was trained on 70,813 clips (113 h), evaluated on 13,389 clips (21.5 h), and tested on 13,420 clips (21.6 h). The model achieved a WER of 0.12 and a character error rate of 0.03 on the test set (Table 13).

9 | Experiences, Challenges, and Lessons Learned

While collecting and curating the text and speech datasets, we gained valuable experience and faced several challenges. In this section, we summarize these experiences and challenges for communities that plan to collect NLP datasets for low-resourced languages. The data collection and curation process can be described in two main stages:

- *Community engagement*: Where communities are identified, and communication with different stakeholders in the community is carried out [56]. Building a robust community is critical to supporting and enabling the rest of the data collection process.
- *Text and voice data collection*: Text and speech datasets are curated and collected with support from the community.

9.1 | Lesson One: Communicate the Value of Data Collection

While building the community, the greatest challenge is determining and communicating the value added to the community. Stakeholders must understand and relate to the benefits of participating in the data collection process. This can especially be challenging when building open source datasets as it can limit some tangible benefits like recognition or use of the datasets. Members of the community need to appreciate the importance of contributing to the process to have a connection to it and to make it sustainable. For example, the value of the datasets as communicated to researchers who are typically interested in access to quality data will be different from the value communicated to cultural institutions that may have more interest in language preservation and advancement of the languages [6]. To enable effective engagement with potential community stakeholders, there needs to be a clear communication strategy that defines the profile of each stakeholder to understand their perspective of the process and how it might be of value to them. The development of this strategy should include representation from different stakeholder groups to guarantee an inclusive approach to communication.

9.2 | Lesson Two: Have an Inclusion Strategy for Community Contributions

Stakeholders in the community are diverse, and inclusion needs to be carefully evaluated while building the community. There is varying demographic representation among communities, and care should be taken to develop a well-represented community. For example, in this project, different age groups were represented in different communities, and this was taken into account to build an age-inclusive speech dataset. The most important question to ask oneself while building a community is "how do I effectively communicate the importance of the process to the communities?" In many cases, this comes down to presenting the case and thinking about how the community's efforts will be rewarded. One solution, especially when it comes to rewards, is to devise a mechanism to reward community members for their effort in the project.

For example, the leading data collectors were awarded gifts to recognize their efforts toward contributing to data collection on the project. In many cases, this can encourage others to participate. To ensure inclusion while building the community,

one has to effectively study what inclusion means from the perspective of the desired datasets. Inclusion can be influenced by some high-level factors such as gender or racial representation or specific factors such as age or industry representation. An inclusion strategy would generally involve understanding the desired use of the final dataset and exploring the different stakeholders who need to contribute to the process and ensuring that there is no intentional exclusion of critical stakeholders. Although this may not entirely solve the challenge, considering inclusion will help develop better language datasets.

9.3 | Lesson Three: Have a Plan for the Text Data Collection

The data collection stage is where most of the challenges were encountered during the project. The main data types collected during the project were text corpus data collection and voice corpus generation based on the text corpus. Text corpus generation is critical to ensure a quality voice dataset; therefore, the first issue is developing a robust text data collection strategy. This involved the development of guidelines that would be used to collect the text datasets. The text data were curated from two main sources: existing text and original creations from the community. The existing text sources can be further categorized by availability: public domain text and private text. The main challenge faced with the inclusion of public sources into the text corpus was ensuring that the text conforms to the specified standards. For example, text data published on CV [46] must meet the specific requirements described earlier.

During this project, many public text datasets needed to be modified before being included in the text corpus. This should be considered while making the data collection plan. One may need to evaluate the effort required to modify the dataset, which can inform which datasets should be considered for addition to the text corpus. Licensing is another challenge to consider while evaluating the adoption of public datasets because licensing varies even in the public domain. For example, the Creative Commons license [57] which is one of the commonest licenses under which public datasets are published, has different versions, which might restrict the usage of the datasets for specific applications. It is critical to ensure the public data license is compatible with the licenses under which the text corpus will be published. For example, the data on CV [46] are published under the CC-0 license [58] which is typically viewed as a "no-strings" attached license, and this may not be compatible with the Creative Commons by Attribution 4.0 (CC by 4.0) [59] license, thus limiting the adoption of such datasets.

9.4 | Lesson Four: Develop Fair and Agreeable Licensing Structures for Text Corpus Creation

The main challenges faced while developing community-curated text datasets are licensing and validation. Existing creations such as storybooks, radio, and television show transcripts, and other literature constitute the most significant portion of the community data. It is expected that linguists and enthusiasts will develop original literature resources and are willing to share these in the public domain to improve accessibility to different language resources. A fair or agreeable arrangement must be developed to use these datasets. The single biggest obstacle to such agreements is finding the appropriate licensing structure to accommodate both the creator's wishes and the data publication requirements. Licenses like CC by 4.0 can enable this by ensuring that the original creators of such works are attributed by the data users while guaranteeing wide use of the data in the public domain. A smaller portion of the community-curated text data consisted of sentences developed by community members.

9.5 | Lesson Five: Have a Robust Validation Guideline

Although the quality of such sentences can be guided by developing standards for creators to follow, the validation scope typically goes beyond the standards. It is common practice that community members might plagiarize existing sources, which is especially difficult to detect during the data validation; therefore, there needs to be a robust validation pipeline to minimize such instances. Such acts can potentially violate the licensing of the plagiarized text sources and affect the overall quality of the dataset. This challenge is sometimes compounded by the desire for individuals to increase the value of the compensation for their efforts in developing these sentences. Although it is critical to ensure fair compensation, a balance should be made to ensure this does not lead to such behavior. A potential solution to this challenge implemented during the project was to use monetary and community recognition-based rewards, which helped to get more genuine contributors.

9.6 | Lesson Six: Ensure Diversity in Voice Data Collection Process

The challenges encountered during the voice data collection stage primarily relate to limitations during community engagement and text data collection. Perhaps the biggest observed challenge was fair representation in the collected voice corpus, which was evaluated based on age and gender characteristics during this project. There must be sufficient diversity (age and gender) while the voice dataset is being developed. Rather than evaluate this at the end of the project, it needs to be monitored continuously to make appropriate changes to promote the desired diversity in the dataset. The community composition from a gender and age perspective will have the most significant influence on the composition of the voice dataset. This is why it is crucial to have a well-defined inclusion plan during the community engagement and building stage. For example, suppose the community building is focused on university and student communities. In that case, it is expected that the age and gender characteristics of such groups will be the most present in the voice dataset. The continuous monitoring and profiling of these characteristics in the voice dataset can provide valuable insights that can influence community engagement activities to ensure fair representation in the final dataset.

9.7 | Lesson Seven: Ensure Clear Text Data Collection Guidelines

The quality of the voice datasets is directly related to that of the text datasets; a poorly developed text corpus will lead to

many challenges during voice data collection. Issues such as incorrect grammar and complex sentences directly affect the readability and the quality of datasets. One common observation during the data collection process was that users would skip specific sentences more than others, which implied that such sentences had a small representation in the dataset. Sometimes, the issue is not with the technical composition of the sentences. Still, the comprehension of the sentence is an issue that is not easy to capture during text validation and is mainly perceived during voice data collection. The text comprehension issue can originate from modifying the text from the public datasets and the evolution of language across different age groups. Therefore, this must be considered as one embarks on voice data collection.

9.8 | Lesson Eight: Provide Value and Guidelines for Voice Validation Process

The most crucial step in collecting the voice datasets is the validation of the collected voices, which involves community members listening to and approving the collected voice clips. The biggest challenge one may encounter during this process is having a low validation rate for the voice data and it is, therefore, essential to ensure that there is sufficient activity in the community toward the voice validation. However, even if this is the case, inherent challenges mainly arise from the demographic composition of the voice contributors and the validators. It is common to have differences in how people speak; in fact, it is of specific importance that this variation is captured during the voice collection. The biggest problem is that the validators may not be accustomed to these voice variations and may give an erroneous evaluation of a voice clip. Although a theoretical solution to this is to have a majority-based vote on voice quality, this majority depends on how many people have listened to a specific voice clip, and this is not a guaranteed scenario. To minimize this challenge, standards must be developed to guide validators on evaluating a voice clip fairly. The standards must be constantly updated to reflect the changes observed during the process.

9.9 | Lesson Nine: Development of Robust Data Collection Standards

One lesson that is a hallmark of the different challenges during the community development and data collection stages is the development of robust standards that can be used during implementation. This is perhaps the most significant challenge one should expect while curating similar datasets because most existing standards and guidelines are specific to a dataset. For instance, the guidelines used to curate English text for the CV [46] platform may not be compatible with languages like Luganda. A person collecting datasets, especially for new languages, should consider existing guidelines as a starting point, evaluate them, and modify them accordingly to customize the new dataset. Generally speaking, a good data collection strategy standard or guideline should include a detailed description and clearly state the environment and assumptions under which it was developed. This can ensure flexibility in its implementation and guide how it can be adapted to different environments. Ideally, for similar projects, the standards should clearly describe interdependencies between different stages of the data collection and how they influence each other. This could minimize the knockdown effect of challenges arising from poorly implemented stages earlier in the dataset curation process.

Although it is naive to expect a near-perfect data collection process, the success of similar projects can be aided by considering the general lessons learned from the implementation of this project. Community is critical to the success of the data collection process, and one needs to develop inclusive communities and have a good communication strategy to encourage engagement in the community. Good community engagement will enable the continuity of the data collection and curation efforts even beyond the lifetime of the specific project. The main goal of the data collection process is to develop good quality and usable datasets for the desired goal, which the development of good standards and guidelines will significantly simplify.

10 | Conclusion

This paper describes our process of creating text and speech datasets for low-resourced languages in East Africa. Working across five languages, we provided the technicalities in building resources for low-resourced languages. The data creation process is not the same across the languages. We have presented different approaches to collecting text and voice datasets for these languages. We introduced several data collection tools that were used to collect the data. These were used to collect and curate a parallel corpus for the five languages. We have also provided results from baseline models for MT, sentiment analysis, topic classification and modeling, and ASR tasks. While collecting and curating the text and speech datasets, we gained valuable experience and faced several challenges. We have discussed our experiences and challenges for communities that plan to collect NLP datasets for low-resourced languages. Future work should focus on understanding the biases that may result in using these datasets, for example, gender bias and intersectionality evaluation for the speech datasets. Investigating of strategies for creating representative datasets regarding dialects and accents and how this may affect the downstream NLP tasks is required.

Acknowledgments

We want to thank the Department of African Languages, Makerere University, for the translation work on the parallel text corpus. We want to thank Buganda Kingdom for partnering with us and for supporting the Luganda monolingual text collection through its agencies. We would also like to acknowledge the various organizations and individuals from which we sourced these data: the Independent News, Jerum Agency. We want to thank all the community contributors in East Africa who have enabled us to achieve voice contributions. This work was carried out with support from Lacuna Fund—No. 1937.70 2021, an initiative co-founded by The Rockefeller Foundation, Google.org, and Canada's International Development Research Centre, Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ), and Mozilla.

Conflicts of Interest

The authors declare no potential conflict of interest.

Data Availability Statement

The data that support the findings of this study are openly available in Multilingual Parallel Text Corpora for East African Language at https://doi.org/10.7910/DVN/BEROE0.

Endnotes

[1]https://github.com/masakhane-io.

[2]https://CommonVoice.mozilla.org/.

[3]https://zindi.africa/competitions/giz-nlp-agricultural-keyword-spotter.

[4]https://lanfrica.com/.

[5]https://mozilla-pontoon.readthedocs.io/en/latest/.

[6]https://commonvoice.mozilla.org/lg/datasets.

[7]https://commonvoice.mozilla.org/sw/datasets.

[8]https://pontoon.mozilla.org.

[9]https://huggingface.co/.

[10][Gensim framework] https://radimrehurek.com/gensim/models/word2vec.html.

[11][Glove Package] https://nlp.stanford.edu/projects/glove/.

[12]https://scikit-learn.org/stable/.

[13]https://github.com/common-voice/CorporaCreator.

[14]https://github.com/coqui-ai/STT.

[15]https://github.com/ftyers/commonvoice-utils.

[16]https://huggingface.co/facebook/wav2vec2-xls-r-300m.

References

1. M. E. Ssentanda and J. Nakayiza, "'Without English There Is No Future': The Case of Language Attitudes and Ideologies in Uganda," in *Sociolinguistics in African Contexts: Perspectives and Challenges*, eds. A. E. Ebongue and E. Hurst (Switzerland AG: Springer International Publishing, 2017), 107–126.

2. L. Martinus and J. Z. Abbott, "A Focus on Neural Machine Translation for African Languages," 2019, arXiv Preprint arXiv:1906.05685.

3. V. Marivate, T. Sefara, V. Chabalala, et al., "Investigating an Approach for Low Resource Language Dataset Creation, Curation and Classification: Setswana and Sepedi," 2020, arXiv Preprint arXiv:2003.04986.

4. N. Wilhelmina, M. Vukosi, M. Tshinondiwa, et al., "Participatory Research for Low-Resourced Machine Translation: A Case Study in African Languages," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, eds. T. Cohn, Y. He, and Y. Liu (Association for Computational Linguistics, 2020), 2144–2160.

5. D. I. Adelani, J. Abbott, G. Neubig, et al., "MasakhaNER: Named Entity Recognition for African Languages," *Transactions of the Association for Computational Linguistics* 9 (2021): 1116–1131.

6. K. Siminyu, J. Abbott, K. Tubọsun, et al., "Consultative Engagement of Stakeholders Toward a Roadmap for African Language Technologies," *Patterns* 4, no. 8 (2023): 100820.

7. P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury, "The State and Fate of Linguistic Diversity and Inclusion in the NLP World," 2020, arXiv Preprint arXiv:2004.09095.

8. D. Amodei, S. Ananthanarayanan, R. Anubhai, et al., "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, eds. M. F. Balcan and K. Q. Weinberger (New York: Proceedings of Machine Learning Research, 2016), 173–182.

9. F. Boyer and J. L. Rouas, "End-to-End Speech Recognition: A Review for the French Language," 2019, arXiv Preprint arXiv:1910.08502.

10. B. C. Youcef, Y. M. Elemine, B. Islam, and B. Farid, *Speech Recognition System Based on OLLO French Corpus by Using MFCCs* (Switzerland AG: Springer, 2017), 326–331.

11. S. Zhou, L. Dong, S. Xu, and B. Xu, *A Comparison of Modeling Units in Sequence-to-Sequence Speech Recognition With the Transformer on Mandarin Chinese* (Switzerland AG: Springer, 2018), 210–220.

12. S. Karita, Y. Kubo, M. A. U. Bacchiani, and L. Jones, "A Comparative Study on Neural Architectures and Training Methods for Japanese Speech Recognition," 2021, arXiv Preprint arXiv:2106.05111.

13. B. F. Dossou and C. C. Emezue, "OkwuGb'e: End-to-End Speech Recognition for Fon and Igbo," 2021, arXiv Preprint arXiv:2103.07762.

14. L. Fund, "Language Domain," 2023, https://lacunafund.org/datasets/language/, Last accessed December 1, 2023.

15. D. Adelani, D. Ruiter, J. O. Alabi, et al., "MENYO-20k: A Multi-Domain English-Yorubá Corpus for Machine Translation and Domain Adaptation," 2021, CoRR, abs/2103.08647.

16. A. Öktem, E. DeLuca, R. Bashizi, E. Paquin, and G. Tang, "Congolese Swahili Machine Translation for Humanitarian Response," 2021, arXiv Preprint arXiv:2103.10734.

17. I. Ezeani, P. Rayson, I. Onyenwe, C. Uchechukwu, and M. Hepple, "Igbo-English Machine Translation: An Evaluation Benchmark," 2020, https://arxiv.org/abs/2004.00648.

18. S. H. Muhammad, D. I. Adelani, S. Ruder, et al., "NaijaSenti: A Nigerian Twitter Sentiment Corpus for Multilingual Sentiment Analysis," 2022, arXiv Preprint arXiv:2201.08277.

19. M. Diallo, C. Fourati, and H. Haddad, "Bambara Language Dataset for Sentiment Analysis," 2021, arXiv Preprint arXiv:2108.02524.

20. S. H. Muhammad, I. Abdulmumin, A. A. Ayele, et al., "Afrisenti: A Twitter Sentiment Analysis Benchmark for African Languages," 2023, arXiv Preprint arXiv:2302.08956.

21. I. Shode, D. I. Adelani, and A. Feldman, "yosm: A New Yoruba Sentiment Corpus for Movie Reviews," 2022, arXiv Preprint arXiv:2204.09711.

22. W. F. Oyewusi, O. Adekanmbi, I. Okoh, et al., "Naijaner: Comprehensive Named Entity Recognition for 5 Nigerian Languages," 2021, arXiv Preprint arXiv:2105.00810.

23. R. Mbuvha, D. I. Adelani, T. Mutavhatsindi, et al., "MphayaNER: Named Entity Recognition for Tshivenda," 2023, arXiv Preprint arXiv:2304.03952.

24. Masakhane, "Masakhane," 2023, https://www.masakhane.io, Last accessed 25 March 2023.

25. I. Orife, J. Kreutzer, B. Sibanda, et al., "Masakhane–Machine Translation for Africa," 2020, arXiv Preprint arXiv:2003.11529.

26. D. I. Adelani, J. O. Alabi, A. Fan, et al., "A Few Thousand Translations Go a Long Way! Leveraging Pre-Trained Models for African News Translation," 2022, arXiv Preprint arXiv:2205.02022.

27. C. C. Emezue and F. P. B. Dossou, "FFR v1. 1: Fon-French Neural Machine Translation," in *Proceedings of the Fourth Widening Natural Language Processing Workshop* (2020), 83–87.

28. O. Ahia and K. Ogueji, "Towards Supervised and Unsupervised Neural Machine Translation Baselines for Nigerian Pidgin," 2020, arXiv Preprint arXiv:2003.12660.

29. D. I. Adelani, G. Neubig, S. Ruder, et al., "MasakhaNER 2.0: Africa-Centric Transfer Learning for Named Entity Recognition," 2022, arXiv Preprint arXiv:2210.12391.

30. AI S, "African Technology Initiate for AI for Social Good," 2019, http://www.sunbird.ai/, Last accessed 09 June 2021.

31. B. Akera, J. Mukiibi, L. S. Naggayi, et al., "Machine Translation for African Languages: Community Creation of Datasets and Models in Uganda," in *3rd Workshop on African Natural Language Processing*, 2022.

32. K. Siminyu, G. Kalipe, D. Orlic, et al., "AI4D—African Language Program," 2021, arXiv Preprint arXiv:2104.02516.

33. J. Mukiibi, B. Claire, and N. N. Joyce, "An English-Luganda Parallel Corpus," *Zenodo*, May 2021, 15, https://doi.org/10.5281/zenodo.4764038.

34. R. Lastrucci, I. Dzingirai, J. Rajab, et al., "Preparing the Vuk'uzenzele and ZA-gov-Multilingual South African Multilingual Corpora," 2023, arXiv Preprint arXiv:2303.03750.

35. B. Wanjawa, L. Wanzare, F. Indede, O. McOnyango, E. Ombui, and L. Muchemi, "Kencorpus: A Kenyan Language Corpus of Swahili, Dholuo and Luhya for Natural Language Processing Tasks," 2022, arXiv Preprint arXiv:2208.12081.

36. J. Mukiibi, A. Katumba, J. Nakatumba-Nabende, A. Hussein, and J. Meyer, "The Makerere Radio Speech Corpus: A Luganda Radio Corpus for Automatic Speech Recognition," in *Proceedings of the Language Resources and Evaluation Conference,* eds. N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis (Marseille, France: European Language Resources Association, 2022), 1945–1954.

37. J. Mukiibi, B. Claire, and N. N. Joyce, "Agriculture Keywords Dataset," *Zenodo*, December 2020, 15, https://doi.org/10.5281/zenodo.4347307.

38. D. Blachon, E. Gauthier, L. Besacier, G. N. Kouarata, M. Adda-Decker, and A. Rialland, "Parallel Speech Collection for Under-Resourced Language Studies Using the LIG-Aikuma Mobile Device App," *Procedia Computer Science* 81 (2016): 61–66.

39. D. M. Eberhard, G. F. Simons, and C. D. Fennig, *Ethnologue: Languages of the World*, 23rd ed. (Dallas, TX: SIL International, 2020).

40. J. Nakayiza and N. Yoneda, "Ganda (JE15)—Descriptive Materials of Morphosyntactic Microvariation in Bantu" (PhD thesis, Tokyo University of Foreign Studies, 2023).

41. "Nkore Language," 2023, https://en-academic.com/dic.nsf/enwiki/11824929, Last accessed 1 December 2023.

42. "Masaba Language," 2023, https://dbpedia.org/page/Masaba_language, Last accessed 1 December 2023.

43. S. Lorenz, "Living With Language. An Exploration of Linguistic Practices and Language Attitudes in Gulu, Northern Uganda" (PhD thesis, Universität zu Köln, 2019).

44. K. Ngugi, W. Okelo-Odongo, and P. Wagacha, "Swahili Text-to-Speech System," *African Journal of Science and Technology* 6, no. 1 (2005): 80–89.

45. C. Babirye, J. Tusubira, J. Mukiibi, J. Nakatumba-Nabende, and A. Katumba, "Sentiment Tagged Parallel Corpus for Luganda and Swahili," Harvard Dataverse 1 2023, https://doi.org/10.7910/DVN/XSGIKR.

46. R. Ardila, M. Branson, K. Davis, et al., "Common Voice: A Massively-Multilingual Speech Corpus," 2019, arXiv Preprint arXiv:1912.06670.

47. A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention Is All You Need," in *Advances in Neural Information Processing Systems 30* (Long Beach, CA: Curran Associates Inc, 2017).

48. M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, et al., "Marian: Fast Neural Machine Translation in C++," 2018, arXiv Preprint arXiv:1804.00344.

49. A. Hyun and L. Soo-Young, "Hierarchical Representation Using NMF," in *International Conference on Neural Information Processing, 159,* eds. M. Lee, A. Hirose, Z.-G. Hou, and R. M. Kil (Berlin, Heidelberg: Springer, 2013), 466–473.

50. P. Cristian and E. B. Mihaela, "Dealing with Data Imbalance in Text Classification," *Procedia Computer Science* 159 (2019): 736–745.

51. P. Zhou, Z. Qi, S. Zheng, J. Xu, H. Bao, and B. Xu, "Text Classification Improved by Integrating Bidirectional LSTM With Two-Dimensional Max Pooling," 2016, arXiv preprint arXiv:1611.06639.

52. J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," 2014, arXiv preprint arXiv:1412.3555.

53. A. Hannun, C. Case, J. Casper, et al., "Deep Speech: Scaling Up End-to-End Speech Recognition," 2014, arXiv Preprint arXiv:1412.5567.

54. K. Heafield, "KenLM: Faster and Smaller Language Model Queries," in *Proceedings of the Sixth Workshop on Statistical Machine Translation* (2011), 187–197.

55. A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised Cross-Lingual Representation Learning for Speech Recognition," 2020, arXiv Preprint arXiv:2006.13979.

56. W. Nekoto, V. Marivate, T. Matsila, et al., "Participatory Research for Low-Resourced Machine Translation: A Case Study in African Languages," 2020, arXiv Preprint arXiv:2010.02353.

57. L. Lessig, "The Creative Commons," *Montana Law Review* 65 (2004): 1.

58. Creative Commons, "CC0: Public Domain Dedication," 2013.

59. M. Debnath, P. Doubrawa, M. Optis, P. Hawbecker, and N. Bodini, "The Creative Commons Attribution 4.0 License," 2021.