American Journal of Artificial Intelligence
2025; 9(2): 167-185
http://www.sciencepublishinggroup.com/j/ajai
doi: 10.11648/j.ajai.20250902.18

Bridging Swahili Communication Gaps: Real-Time Audio-to-Text Sentiment Analysis via Pre-trained NLP

Kevin Obote[1, *], Benjamin Kikwai[2], Kennedy Senagi[1], Joyce Njiiri[3], John Olukuru[1], Joseph Sevilla[1]

[1]Strathmore Institute of Mathematical Science, Strathmore University, Nairobi, Kenya
[2]Department of Mathematics and Statistics, Machakos University, Machakos, Kenya
[3]Department of Languages and Lingustics, Machakos University, Machakos, Kenya

Email address:

kevin.obote@strathmore.edu (Kevin Obote), kikwaib@mksu.ac.ke (Benjamin Kikwai), ksenagi@strathmore.edu (Kennedy Senagi), joycenjiiri@mksu.ac.ke (Joyce Njiiri), jolukuru@strathmore.edu (John Olukuru), joe@strathmore.edu (Joseph Sevilla)
[*]Corresponding author

To cite this article:

Kevin Obote, Benjamin Kikwai, Kennedy Senagi, Joyce Njiiri, John Olukuru, Joseph Sevilla. (2025). Bridging Swahili Communication Gaps: Real-Time Audio-to-Text Sentiment Analysis via Pre-trained NLP. *American Journal of Artificial Intelligence, 9*(2), 167-185. https://doi.org/10.11648/j.ajai.20250902.18

Received: 4 August 2025; **Accepted:** 18 August 2025; **Published:** 25 September 2025

Abstract: The global proliferation of digital communication highlights a critical gap in language technologies for digitally under-represented languages, particularly Kiswahili, a language spoken by over 100 million people. While significant advancements have been made in natural language processing (NLP) for high-resource languages like English, a persistent challenge remains in creating robust computational systems for low-resource linguistic contexts. This study addresses this challenge by presenting a novel, end-to-end Kiswahili audio processing pipeline that unifies three core capabilities; real-time speech recognition, sentiment analysis, and text summarization. The system's novelty lies in its strategic leverage of state-of-the-art, pre-trained machine learning models, including Wav2vec2, DistilBERT, and T5, demonstrating a viable approach to bridging the digital communication gap for Kiswahili in real-world applications. Our methodology involved a rigorous evaluation of the integrated system using the Mozilla Common Voice Corpus. The results revealed key insights and promising performance metrics. The speech recognition component, a foundational element of the pipeline, achieved an exceptionally low Word Error Rate (WER) of 0.3329 with the Wav2vec2 model, highlighting its capacity for accurate transcription in a low-resource setting. This is a significant finding, as it suggests that models specifically fine-tuned for such environments can overcome the challenges of data scarcity and linguistic diversity. The summarization component also demonstrated strong capabilities, yielding a ROUGE-L score of 0.6622, which indicates robust semantic and structural alignment with reference texts. While the sentiment analysis revealed a notable data imbalance with a predominance of negative samples, the model achieved a 60% accuracy, demonstrating its potential for further refinement. These findings underscore both the immense potential and the inherent limitations of applying pre-trained models to a low-resource language like Kiswahili. They provide a compelling proof of concept for the technical feasibility of Kiswahili audio processing and emphasize the critical need for continued investment in dataset expansion and model optimization. The study concludes that this work establishes a foundational groundwork for continued research and the subsequent development of advanced NLP tools specifically tailored for Kiswahili-speaking populations, ultimately aiming to improve access to education, healthcare, and information services, and to foster greater digital inclusion throughout East Africa.

Keywords: Automatic Speech Recognition, Natural Language Processing, Kiswahili

1. Introduction

The linguistic diversity of Africa, with over 2,000 languages spoken across the continent, presents unique challenges. While this diversity is a cultural asset, it complicates efforts to create inclusive communication systems, especially in contexts such as education, healthcare,

and governance [1]. These communication barriers can hinder effective engagement with diverse communities, highlighting the need for targeted strategies to accommodate Africa's rich linguistic tapestry while promoting mutual understanding and cooperation among its various language-speaking populations. Many indigenous languages are not represented in formal education, governance, or digital platforms, leaving speakers marginalized [2]. The under representation of African languages in natural language processing tools (NLP) exacerbates this problem, limiting the accessibility of digital technologies for millions of people [3]. One of the most significant milestones in African communication has been the rise of Kiswahili as a unifying regional language. Spoken by over 100 million people, Kiswahili acts as a lingua franca across East Africa, facilitating cross-border trade, diplomacy, and cultural exchange [4]. Its adoption by the African Union as one of its official languages underscores its potential as a pan-African language [5]. However, despite its prominence, Kiswahili and other indigenous languages remain underutilized in digital communication technologies, reflecting broader systemic biases favoring European languages in technological and educational frameworks [6]. To address these challenges, there is an urgent need for localized Artificial Intelligence (AI) and NLP solutions that prioritize African languages. Such technologies could enhance access to information, promote social inclusion, and empower communities by enabling communication in native tongues [7]. Efforts to integrate African languages into digital spaces would not only preserve cultural heritage, but would also bridge gaps in education, governance, and economic participation [8].

1.1. Linguistic Diversity and Socio-Economic Disparities

East Africa is home to a rich indigenous languages, with more than 100 such languages spoken throughout the region [9]. While Kiswahili has been widely embraced as a common language, the dominance of English in government institutions, business, and international relations often leaves non-native English speakers at a disadvantage [10]. This division creates socio-economic disparities, particularly for rural populations or those in marginalized communities who may not be proficient in English [11]. Furthermore, even though Kiswahili is widely spoken in East Africa, it is often viewed as less prestigious compared to English, especially in education and business settings [12]. This perception can perpetuate cycles of inequality, where those with better access to English-language education have more opportunities for higher-paying jobs, political representation, and social mobility. The linguistic divide, it may seem minor on the surface, has significant implications for the development of East Africa. It contributes to exclusion in decision-making processes, limits access to crucial information, and creates barriers to full participation in the economy, especially in sectors like finance, technology, and governance [13].

1.2. Role of Natural Language Processing (NLP) in Addressing Linguistic Barriers

Natural Language Processing (NLP) technologies have emerged as a promising tool for breaking down linguistic barriers in the digital space. NLP is a field of artificial intelligence that focuses on the interaction between computers and human languages, allowing for the automation of tasks such as translation, speech recognition, and text analysis [14]. In multilingualism context, NLP technologies have the potential to empower communities by enabling better communication in various languages,including Kiswahili and other indigenous languages [15]. AI-driven translation services like Google Translate and speech-to-text applications can significantly enhance cross-border communication, making services and information more accessible to a broader population [16]. In education, NLP can help breakdown language barriers between students and teachers, especially in regions where English is the primary language of instruction but may not be the first language of learners. NLP tools can also aid in the development of local content and help disseminate information in a more inclusive manner, promoting literacy, healthcare,and government services [17].

1.3. Challenges in Mainstream NLP Tools and Digital Exclusion

Despite the potential of NLP technologies, mainstream NLP tools are often not designed to adequately support African languages, including Kiswahili [18]. For instance, popular tools like Google Translate offer limited support for Kiswahili compared to widely spoken languages like Spanish, French, or Chinese [19]. Even though Kiswahili is spoken by millions, the digital resources available for it are often inaccurate or insufficient for advanced applications, such as speech-to-text or real-time translation services [20]. Moreover, there is a lack of robust NLP resources tailored to the many regional languages in East Africa. African indigenous languages often face challenges such as a lack of standardization in orthography, limited digital content, and lower representation in machine learning datasets. These issues prevent the development of high-quality NLP systems that can handle the nuances of African indigenous languages effectively, further contributing to digital exclusion [21]. This gap in representation means that many East Africans who are not proficient in English are excluded from accessing the full potential of digital technologies. It also limits their ability to participate in the growing digital economy, where services like e-commerce, mobile banking, and digital government services are increasingly conducted in English or poorly supported African languages [22].

1.4. Diverse Linguistic Landscape

Kenya's linguistic diversity is a microcosm of the broader East African region. The country is home to over 40 indigenous languages, including Gikuyu, Dholuo, Kamba,

Luhya, and others, which reflect the rich cultural heritage of its people [23]. Alongside these, Kiswahili functions as the lingua franca, unifying the country and acting as a bridge between diverse communities [24]. English, inherited from the colonial era, has taken on the role of the official language in many African countries, dominating sectors such as education, governance, and business [25]. his dual-language system, where both Kiswahili and English are prevalent, provides a practical framework for communication in a multilingual society, allowing for broader access to information and resources. However, it also presents significant challenges, as this framework often sidelines indigenous languages, creating barriers for communities that primarily communicate in these local languages [26]. Non English speakers, may struggle to access quality education, government services, and economic opportunities, perpetuating socio-economic inequalities [27].

1.5. Challenges of Language Marginalization

Language marginalization has significant implications for Kenya's socio-economic development. For many citizens, limited proficiency in English or Kiswahili means restricted access to public services, legal systems, and educational resources [28]. This can lead to a sense of exclusion and reduce participation in national development. In the education sector, the emphasis on English as the medium of instruction can alienate children from communities where indigenous languages are dominant. Learners may struggle to grasp concepts in a language they are not fluent and proficient in, leading to poor performance and high dropout rates [29]. In governance and politics, language barriers can prevent citizens from fully engaging in decision-making processes, especially at the grassroots level. Similarly, in business, English-centric communication often excludes small-scale entrepreneurs and traders operating in informal markets, where indigenous languages are more prevalent [30]. This limits their ability to expand their operations, access credit, or participate in formal trade networks.

Research Objectives

This study aims to develop an innovative real-time audio-to-text system with integrated Natural Language Processing (NLP) capabilities, including summarization, paraphrasing, and sentiment analysis. With focus on Kiswahili language. The research objectives are:

1. *To enhance Linguistic Accessibility and Engagement:* To facilitate real-time transcription in Kiswahili and English, in order to improve comprehension, active participation, and information accessibility for diverse linguistic groups in educational, business, and public forums, especially for non English speakers. This will mitigate socio-economic barriers and promote social equity.
2. *Promote Inclusivity Across Key Sectors:* To provide tools that enhance communication in education through multilingual interactions, ensure transparency and accessibility in governance by transcribing public discourse, and foster seamless business communication among diverse linguistic groups, reducing misunderstandings and building trust.
3. *Support United Nations Sustainable Development Goals (SDGs):* This project directly contributes to:
    1) *SDG 4 (Quality Education):*Improving access to educational content in Kiswahili and in other indigenous languages.
    2) *SDG 10 (Reduced Inequalities):*Addressing language-based barriers to opportunities.
    3) *SDG 9 (Industry, Innovation, and Infrastructure):* Promote technological innovation adapted to local contexts.

By achieving these objectives, this research seeks to transform linguistic diversity into a strategic asset, fostering inclusivity, bridging communication gaps, and empowering individuals. The focus on localized NLP solutions will drive equitable technological growth and digital inclusion in Kenya and the broader East African region.

2. Review of Literature

The integration of real-time audio-to-text and multilingual sentiment analysis is profoundly transforming communication, learning, and accessibility. This review of literature examines current advancements, applications, and challenges of these technologies, focusing on their implications for language learning and communication in Kenya and East Africa, where Kiswahili and English langauges are widely spoken. By synthesizing the state-of-the-art in real-time speech recognition, sentiment analysis, and multilingual natural language processing (NLP), this review provides an in-depth exploration of how these technologies can bridge linguistic divides and enhance social, educational, and economic outcomes.

2.1. Real Time Audio to Text Systems

Real-time audio-to-text systems are designed to convert spoken language to written text almost instantaneously, facilitating communication in diverse settings such as education, healthcare, and governance. Recent advancements in Automatic Speech Recognition (ASR) technology, particularly through the use of deep learning models and Artificial Intelligence, have significantly improved the accuracy and reliability of these systems, extending their presence across various research and business domains [32]. The ability to transcribe speech with minimal latency is crucial for applications such as live captions for media broadcasts and interactive voice-based assistants.

However, the Kenyan multilingual nature, where code-switching translanguaging among English, Kiswahili, and various indigenous languages is prevalent, poses unique challenges for Automatic Speech Recognition (ASR) systems. These challenges necessitate the development of high-quality datasets that capture instances of code-switching

and mixed-language speech(translanguaging), which are particularly limited in East Africa. The need for robust ASR models capable of processing languages with diverse phonetic characteristics and minimal digital resources has been emphasized in recent research [37, 38].

The benchmarking of model performance is a critical component in the continuous improvement of Automatic Speech Recognition (ASR) systems, particularly in multilingual contexts. Recent studies illustrate significant advancements achieved in specific languages. For instance, research focusing on multilingual ASR for various languages indicates a notable endeavor to refine ASR technology for low-resource languages, though specific quantitative figures for languages like Kinyarwanda, Swahili, and Luganda are not well-documented in the current literature [79].

Further demonstrating the progress in this field, specialized models can yield improved performance over more general systems, showcasing the importance of targeted training. Recent models have shown performance improvements; however, exact reported Word Error Rates (WERs) for Swahili, such as the 10.2% from the Scribe v1 model on the Few-Language Evaluation of Universal Representations of Speech (FLEURS) benchmark or the 5.09% from a fine-tuned Whisper Tiny model [80, 81].

The overall landscape is characterized by an ongoing concerted effort to lower error rates across different languages, thereby establishing rigorous benchmarks for future performance evaluation. The remarkable figures reflect advancements made within the scope of current research, with implications suggesting a pathway toward further efficiency and effectiveness in ASR applications. Thus, these developments signify not only an enhancement in technological capability but also an ongoing commitment to improving accessibility and usability in ASR across diverse linguistic landscapes.

Studies have shown that multilingual ASR systems, which leverage transfer learning techniques, demonstrate effective performance across numerous languages. For example, [39] demonstrated that their system, Language-Agnostic Multilingual ASR and Speech Translation model Using neural transducers(LAMASSU) , significantly reduces model size while outperforming both monolingual ASR systems and bilingual translation models, indicating the potential for effective multilingual handling in ASR systems.

Additionally, research highlights the importance of large-scale multilingual corpora for training these models. By fine-tuning on language-specific datasets, systems can achieve high accuracy when recognizing speech in both English and Kiswahili [39, 40] also discussed the release of Common Voice Speech-to-Text translation corpus (CoVoST 2), which is a large-scale multilingual speech-to-text corpus, thus fostering further research in the area of low-resource language pairs [39]. Other studies have similarly pointed to improvements seen in ASR systems when employing multilingual strategies. [41] proposed a shared-hidden-layer architecture that has shown success in low-resource speech recognition.

Moreover, practical implementations of multilingual ASR systems have the potential to significantly enhance communication and accessibility in linguistically diverse regions. Recent literature suggests that employing dual-decoder architectures that tightly integrate ASR and translation tasks could improve the overall effectiveness of multilingual systems, as seen in the work of [42]. The flexibility of multilingual neural networks to adapt to various languages is supported by empirical data from extensive experiments [38], showcasing their resilience even in low-resource environments.

Beyond mere transcription, the integration of Natural Language Processing (NLP) capabilities, such as summarization, paraphrasing, and sentiment analysis, is crucial for deeper linguistic understanding. A systematic review of multilingual sentiment analysis highlights common languages supported, effective pre-processing techniques (like tokenization, normalization, N-gram, and machine translation), and hybrid classification methods, English remains the most studied language in this field [31]. Deep learning has emerged as a powerful technique for sentiment analysis, learning multiple layers of data representations to produce state-of-the-art prediction results [34]. Recent research compares different sentiment analysis techniques, such as Valence Aware Dictionary and sEntiment Reasoner(VADER) (Bag of Words approach) and the RoBERTa model, demonstrating varying degrees of accuracy and underscoring the importance of model selection for deciphering public opinion and emotional cues in user-generated content [33]. Despite promising results, particularly in domains like financial forecasting, challenges remain, including the requirement for large datasets and the mitigation of biases in models [36]. Moreover, NLP techniques are vital for analyzing textual feedback in educational contexts, assisting in identifying areas for improvement through sentiment annotations, text summarization, and topic modeling, while also addressing context-based challenges like sarcasm, domain-specific language, and ambiguity [32]. The application of NLP extends to exploring mental health insights from social media data, further illustrating its capacity to handle complex, real-world linguistic data [35].

2.2. Sentiment Analysis and Emotion Detection

Sentiment analysis plays a crucial role in understanding the emotional tone of spoken or written content, involving the identification and categorization of emotions expressed in text or speech [43]. As a popular area within Natural Language Processing (NLP), it effectively mines implicit emotions in text, aiding enterprises and organizations in making effective decisions [43]. The integration of sentiment analysis with audio-to-text systems allows for a more nuanced understanding of communication, providing insights into user engagement and emotional responses, aligning with recent trends in deep learning-based NLP [44].

In the context of African languages, significant strides have been made in establishing quantitative benchmarks for sentiment analysis. A notable study by Mabokela [82]

investigated sentiment analysis across five Southern African languages; Sepedi, Sesotho, Setswana, isiXhosa, and isiZulu, and reported an average weighted F1 score exceeding 63%. This indicates substantial potential within these languages, although challenges related to limited resources still persist [83].

Additionally, research focused on Swahili sentiment analysis, which utilized a pre-trained BERT-based model on a challenging dataset, yielded an accuracy of approximately 66.7% [84]. These findings highlight the impact of pre-trained language models on enhancing sentiment classification accuracy, particularly for languages with limited datasets. The benchmarks from these studies illustrate both the advancements and the obstacles present in the realm of sentiment analysis for African languages, reinforcing the necessity for dedicated resources and model fine-tuning to improve performance in specific languages and dialects [83].

These efforts illuminate the current state of sentiment analysis in under-resourced languages and serve as a call to action for continued research and development, aiming to bridge the data availability gap and enhance the performance of models designed for diverse linguistic contexts [85, 86].

Recent advancements in machine learning and NLP have led to the development of sophisticated sentiment analysis models. Transfer learning has emerged as a key technique, utilizing existing knowledge to solve problems across different domains and producing state-of-the-art prediction results in sentiment analysis [43]. For instance, in the analysis of interview transcripts, NLP techniques are employed to generate sentiment analysis and perform quantitative techniques, bringing objectivity to analyses that often lack it and helping researchers derive insights by finding patterns [45]. Furthermore, for multimodal sentiment analysis, which extends beyond language to include other modalities like gestures and voice, novel models such as the Tensor Fusion Network have been introduced. This network learns both intra-modality and inter-modality dynamics end-to-end, outperforming state-of-the-art approaches for both multimodal and unimodal sentiment analysis in experiments [48].

The application of sentiment analysis extends to real-time scenarios. A unique method for automatic real-time sentiment analysis of online shopping product reviews employs a combination of NLP strategies and machine learning algorithms (e.g., Support Vector Machines (SVM), Naive Bayes, and Random Forest). This approach aims for a high level of accuracy in categorizing reviews as positive, negative, or neutral, ensuring real-time processing and quick responses to users, thereby enhancing understanding of customer thoughts and preferences [46]. Similarly, for integrated real-time big data stream sentiment analysis, systems like the Social Media Data Stream Sentiment Analysis Service (SMDSSAS) have been developed. These systems perform multiple phases of sentiment analysis on massive volumes of unstructured text streams from sources like social media. Experiments with Deterministic and Probabilistic sentiment models on Twitter data during the 2016 Presidential election demonstrated an average accuracy of 81% for the Deterministic model and 80% for the Probabilistic model, representing a 1% to 22% improvement over existing literature in predicting public opinion [47].

In the context of Kenya and East Africa, where multilingualism is prevalent, the ability to analyze sentiment in both Kiswahili and English is of paramount importance. This capability can provide valuable insights into user engagement, emotional responses, and interpersonal dynamics. For example, in customer service applications, real-time sentiment analysis can help businesses gauge customer satisfaction, allowing for immediate intervention and issue resolution [46, 47]. The continuous growth of data presents both opportunities and challenges for sentiment analysis, particularly in diverse linguistic environments [43].

2.3. Challenges in Multilingual Speech Recognition and Sentiment Analysis

Despite the advancements in sentiment analysis technologies, several challenges remain, particularly in multilingual settings. One major challenge is the scarcity of high-quality, annotated audio-text datasets that reflect the diversity of speech patterns and dialects found in different regions. For instance, research on categorizing sentiment polarities in social network data using Convolutional Neural Networks (CNN) highlights the need for efficient methods to extract useful information from vast user data, with CNN models achieving high accuracy levels of up to 94.47% for Phone, 95.4% for Laptop, and 94% for TV review datasets in specific contexts [49], however, in regions like East Africa, the lack of publicly available datasets for languages such as Kiswahili poses a significant obstacle to the development of robust sentiment analysis and ASR systems.

Another prominent challenge is the issue of code-switching and translanguaging, multilingual societies, like Kenya. Code-switching complicates the development of ASR systems because the models must be able to detect and process multiple languages within a single conversation. Current ASR systems often struggle with accurately transcribing code-switched speech, as they are typically trained on single-language datasets [37]. To address this, researchers have explored approaches such as incorporating language identification modules into ASR systems, allowing them to switch between languages dynamically based on detected speech patterns.

Moreover, cultural and contextual nuances in language present additional challenges for sentiment detection systems. These systems must be trained to recognize the emotional context of specific words and phrases within a particular cultural framework, which is particularly important in diverse regions like East Africa. For example, studies on context-dependent sentiment analysis in user-generated videos, using LSTM-based models, have shown 5-10% performance improvement over state-of-the-art methods by enabling utterances to capture contextual information [53]. The limited availability of labeled sentiment data in languages such as Kiswahili further hinders the development of effective sentiment analysis systems tailored to local contexts. Research

on sentiment analysis for Dravidian languages in code-mixed text, for instance, underscores the scarcity of research and datasets for code-mixed scenarios in low-resource languages, highlighting the need for dedicated platforms to investigate such problems [54]. While sentiment-based NLP can boost accuracy in specific applications, such as delirium identification in medical informatics (achieving an accuracy of 0.807 and AUC of 0.930 with NLP [50]), these advancements often rely on well-resourced languages and domains [51]. Furthermore, assessing the agreement between human ratings and lexicon-based sentiment ratings, particularly for open-ended responses, reveals that correlations can be moderate to slight, indicating the complexity of aligning automated sentiment with nuanced human judgment [52].

2.4. Educational and Social Implications

The integration of real-time audio-to-text and sentiment analysis technologies has profound implications for education and social communication in multilingual societies. In educational settings, these technologies can significantly enhance language learning by providing immediate feedback on language skills and emotional engagement. The development of concept-level knowledge bases for sentiment analysis, which automatically associate multiword expressions with emotion labels and polarity values, offers a sophisticated approach to understanding emotional nuances in language [55]. Real-time transcriptions of lectures and discussions can improve accessibility for students with hearing impairments or those who speak different languages. Furthermore, sentiment analysis can provide insights into how students are emotionally responding to the learning environment, allowing educators to adjust their methods and improve student outcomes. Research into applying transfer learning to sentiment analysis in social media has shown that this approach can lead to better performance results with a limited amount of data, a crucial factor for developing effective educational tools where comprehensive datasets might be scarce [56].

In social communication, sentiment analysis can help bridge communication gaps in various contexts, from government services to healthcare. For example, in public health campaigns, Artificial Intelligence-enabled analysis of public attitudes on platforms like Facebook and Twitter can ensure messages effectively reach diverse populations, allowing for better engagement and understanding. An observational study on COVID-19 vaccine sentiments in the UK and US revealed overall averaged positive sentiments of 58% in the UK and 56% in the US, with negative sentiments at 22% and 24% respectively, demonstrating the ability to track public concerns and inform targeted policy interventions [57]. In governance, these technologies can enhance citizen engagement by providing insights into public sentiment and concerns, as sentiment analysis is a process crucial for gathering and analyzing people's opinions to aid decision-making for corporations, governments, and individuals [58].

The integration of real-time audio-to-text and multilingual sentiment analysis technologies presents significant opportunities for enhancing communication, learning, and accessibility, particularly in multilingual contexts. While challenges related to dataset availability, code-switching, and cultural nuances remain, advancements in these technologies continue to push the boundaries of what is possible. As these technologies evolve, their potential to transform education, communication, and social interaction in East Africa and beyond becomes increasingly evident. The ongoing development of robust, multilingual solutions will be crucial in bridging linguistic divides, fostering inclusivity, and driving innovation across sectors.

3. Methodology

This methodology outlines a structured approach for the development of a Kiswahili speech analysis system capable of real-time audio transcription, sentiment analysis, and text summarization. The methodology is divided into key stages: data understanding, data cleaning and preprocessing, exploratory data analysis, feature engineering, machine learning modeling, performance evaluation, model optimization, deployment, and validation strategies [74]. This detailed approach ensures scientific rigor, reproducibility, and effectiveness in developing a multilingual language processing tool for Kiswahili language.

3.1. Data Understanding

The dataset used for this research is the Mozilla Common Voice Corpus 11.0 for Swahili (sw). This corpus is an open-source dataset developed to facilitate automatic speech recognition (ASR) research in underrepresented languages. It was released on September 21, 2022, and provides approximately 14.76 GB of Swahili audio data with about 753 written Swahili sentences for transcription tasksFoundation (2022). The dataset was constructed through crowdsourcing, with volunteers recording Swahili speech samples in response to open-ended speech prompts. Once the recordings were made, other contributors verified the quality of the audio and the accuracy of the transcriptions. To ensure privacy, Mozilla removed any personally identifiable information (PII) from the recordings before they were added to the public dataset. Although the majority of the data comes from East Africa, the open crowdsourcing model allowed for contributions from Swahili speakers around the globe [59]. This table provides a summary of the contributor demographics for the Swahili dataset. The data reveals a notable age imbalance, with nearly half (44%) of contributors falling in the 20-29 age range. This suggests the dataset may be primarily representative of a younger demographic, which could influence the model's performance on speech from other age groups. For a detailed breakdown, please refer to Table 1.

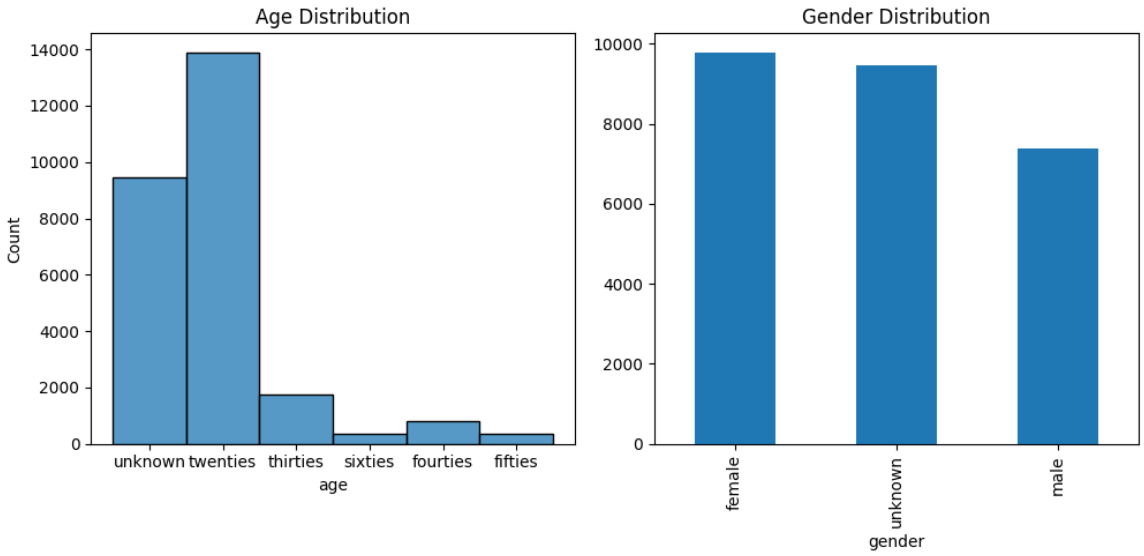

Figure 1. *Distribution of Age and Gender of Dataset Contributors.*

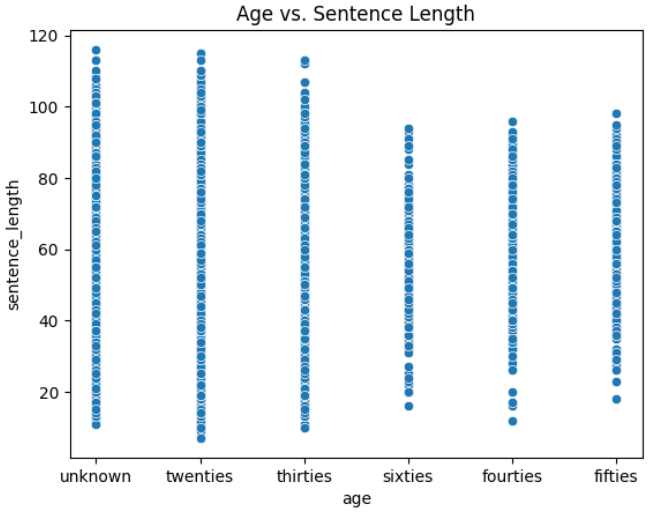

Figure 2. *Bivariate Analysis of Speaker Age and Sentence Length.*

Table 1. *Demographic Breakdown of the Swahili Common Voice Corpus.*

Category	Percentage
Age Groups	
20 - 29	44%
30 - 39	14%
50 - 59	6%
40 - 49	5%
60 - 69	1%
Under 20	0%
Did not disclose age	31%
Gender	
Male	38%
Female or Unspecified	62%

3.2. Data Cleaning and Preprocessing

In order to ensure the accuracy, reliability and efficiency of the machine learning models, the raw data set was subjected to a comprehensive data cleaning and pre-processing pipeline. The primary objective of this step was to remove noise, reduce irrelevant data, and standardize the dataset to facilitate effective model training. The preprocessing involved several stages, each addressing different aspects of the data, including audio trimming, noise reduction, text normalization, and metadata handling.

3.2.1. Silence Trimming

The first step in audio pre-processing was the removal of silent segments from the beginning and end of each audio clip. To accomplish this, the *librosa.effects.trim()* function was utilized. This function automatically detects portions of the audio where the volume falls below a predefined threshold, which generally indicates silence, and trims these sections. This was applied to every audio file to ensure that only the relevant parts of the audio were kept.

The presence of silence in the initial or final segments of audio files offers no useful information for training speech recognition models. These silent segments increase the total duration of the audio clips, which in turn leads to longer processing times and higher computational costs. Furthermore, retaining silence could introduce unnecessary complexity and increase the likelihood of overfitting, as the model might learn patterns unrelated to speech. By trimming the silence, we not only reduced the computational burden but also ensured that the model would focus solely on the speech content, enhancing training efficiency and the generalization capabilities of the final model [60].

Mathematically, the trimming process can be expressed as follows:

$$\text{Trimmed Audio} = \{\text{Audio}[t] \mid \text{Energy}(\text{Audio}[t]) > \text{threshold}\} \tag{1}$$

where:

1. Audio$[t]$ represents the audio signal at time t,
2. Energy$(\text{Audio}[t]) = |\text{Audio}[t]|^2$ is the energy of the audio at time t,
3. threshold is the predefined energy threshold below which the audio is considered silent.

This formula indicates that only the audio segments with energy above the threshold are kept, effectively removing silent parts from the beginning and end of the audio.

3.2.2. Noise Reduction

Next, background noise was reduced to improve the quality of the speech signal. This was achieved using a spectral gating algorithm implemented via the *pydub* library [61]. This algorithm analyzes the frequency spectrum of the audio and suppresses components that fall below a specific noise threshold. This effectively reduces low-level background noise while maintaining the integrity of the speech signal.

Background noise, such as environmental sounds, can significantly hinder the performance of Automatic Speech Recognition (ASR) systems. It creates distortions that can confuse the model during training, leading to poor transcription accuracy. By applying noise reduction, we ensured that the speech signal was clearer and more distinct, making it easier for the model to accurately transcribe and learn from the data. Cleaner audio signals also allow for better generalization across different speakers and environments, ultimately leading to a more efficient model.

The algorithm works by first calculating the noise floor, which represents the background noise level at the beginning of the audio file. The noise threshold is then determined based on this value. Audio components with absolute values below this threshold are suppressed. This process can be described by the following formula:

$$S_{\text{clean}}(f,t) = \begin{cases} S(f,t) & \text{if } |S(f,t)| > \text{threshold}, \\ 0 & \text{if } |S(f,t)| \leq \text{threshold}, \end{cases} \tag{2}$$

Where:

$$\text{threshold} = 2 \times \text{noise\_floor} \tag{3}$$

The noise floor is calculated by taking the mean of the absolute values of the first 1/10th portion of the audio signal, expressed as:

$$\text{noise\_floor} = \frac{1}{n} \sum_{i=1}^{n} |S(f,t)_i| \tag{4}$$

Where:

1. n represents the number of samples in the first 1/10th of the audio signal.
2. $|S(f,t)_i|$ represents the absolute value of the i-th component in the spectrogram.

After determining the noise floor, the noise threshold is set as twice this value, and the audio components with absolute values below the threshold are suppressed, thus achieving noise reduction.

3.2.3. Text Normalization

Text normalization is a critical step in reducing data complexity and ensuring that the model focuses on meaningful patterns in the data. By standardizing the text, to ensure that irrelevant differences, such as capitalization and punctuation, do not interfere with the model's learning process. Furthermore, by removing stopwords, to reduce the noise in the data and allow the model to concentrate on the key components of the text that contribute to its semantic meaning [62]. This leads to a more efficient training process and a model that can generalize better to unseen data.

3.2.4. Handling Missing Values and Metadata Validation

Another important step in the preprocessing pipeline is handling missing metadata.It is always important to handle accordingly any missing metadata Incomplete metadata can introduce bias or inconsistencies into the training process. If records with missing values were discarded, it could lead to a loss of valuable data, especially in a dataset with a large number of contributors. Imputing missing values with a neutral placeholder, such as "unknown," ensures that all data points are represented consistently [62]. This approach helps maintain the integrity of the dataset and prevents bias from affecting the model, particularly when demographic features such as age and gender are incorporated into model predictions. In addition, the uniformity introduced by the placeholders supports better model performance and ensures that no data is unjustifiably excluded from the training process.

3.3. Exploratory Data Analysis (EDA)

3.3.1. Univariate Analysis

Univariate analysis was used to examine the distribution of individual variables, such as age and gender, in isolation. Understanding the distribution of each feature is critical for identifying any potential data imbalances or biases that could affect model performance. Gender distribution: For categorical data such as gender, a bar chart was used to visualize the count of male, female, and unspecified contributors. This visualization helped assess the balance between different gender categories and ensured that the dataset did not exhibit any gender-related biases. Univariate analysis provides insights into the basic structure of the dataset and allows us to detect skewness, outliers, or imbalances in the distribution of data.

3.3.2. Bivariate Analysis

Bivariate analysis was performed to explore relationships between pairs of variables, particularly to understand how characteristics such as the age of the speaker and the complexity of the sentence interact with each other. These analyses help identify potential correlations or dependencies that could be useful for feature engineering.Scatter plots were used to examine the relationship between speaker age and sentence complexity (measured by sentence length or number of words). By plotting age on one axis and sentence length on the other, it was possible to read whether old or younger speakers tend to produce longer or shorter sentences. This analysis is particularly useful for identifying patterns or trends, such as whether certain age groups use more complex language in their speech.

3.4. Feature Engineering

Feature engineering was a critical stage in the preparation of the Swahili speech dataset, as it transformed raw audio data into meaningful representations suitable for machine learning tasks. This process involved a series of carefully selected techniques designed to improve the quality, relevance, and diversity of the features used for model training. The goal was to extract informative elements from the speech signals while reducing noise and irrelevant variations that could affect the performance of the model.

3.4.1. Audio Preprocessing and Quality Refinement

The first step in the feature engineering process focused on audio preprocessing to ensure high-quality data by removing unwanted elements from the recordings. Silence trimming was applied using the *librosa.effects.trim()* function to remove noninformative silent segments from the beginning and end of each audio recording. Silence can extend the duration of the audio without contributing useful information, increasing computational costs and risking overfitting during model training. By trimming silence, the audio clips were made more concise, highlighting the content of the speech and improving data efficiency. Noise Reduction was performed using a spectral gating algorithm implemented with *pydub*. This method reduced background noise by identifying a noise threshold and suppressing low-energy frequency components that were likely unrelated to speech. Clear audio is crucial in speech datasets, as noise can obscure the phonetic details required for tasks such as automatic speech recognition (ASR). By reducing noise, the clarity of the speech signals was enhanced, leading to more accurate feature extraction and downstream performance improvements.

3.4.2. Acoustic Feature Extraction

After preprocessing, acoustic features were extracted to create structured representations of the speech signals. These features captured both spectral and temporal properties of the audio, making them highly effective for speech-related tasks.Mel-Frequency Cepstral Coefficients (MFCCs) were calculated to represent the frequency content of the audio signal. MFCCs are widely used in speech recognition because they provide a compact summary of the spectral envelope, focusing on the frequencies most relevant to human speech perception. The MFCCs were computed using a windowed Fast Fourier Transform (FFT) combined with a Mel-scale filter bank, effectively capturing the phonetic structure of the Swahili speech data. Mel spectrograms were also generated to visualize the frequency distribution of the audio over time. A Mel spectrogram uses a Short-Time Fourier Transform (STFT) to map the audio signal into a two-dimensional time-frequency representation with the frequency axis converted to the Mel scale. This visualization captures both the energy distribution and temporal variations of speech, making it useful for tasks where prosodic elements such as intonation and stress patterns play a role.

Both MFCCs and Mel spectrograms were chosen because they provide a balance between dimensionality reduction and information preservation. These features emphasize speech content while minimizing unnecessary complexity, making them ideal for machine learning applications.

3.4.3. Data Augmentation for Diversity and Robustness

To further improve the dataset's diversity and generalization capacity, data augmentation techniques were applied. Augmentation introduces controlled variations into the dataset, helping the model to learn to generalize across different recording styles and conditions. Pitch Shifting was used to modify the pitch of the audio signal without altering the speech content. Shifting the pitch to the highest and lowest position simulates variations in vocal tone between different speakers. Time stretching was another augmentation method that was applied, which altered the speed of the audio while maintaining the original pitch. This technique helped create variability in speaking rates, making the dataset more representative of natural speech patterns. These enhancement techniques generated additional data points from the original recordings, improving the model's ability to generalize between speakers with different vocal characteristics and speaking speeds.

3.4.4. Visualization of Audio Data and Feature Representations

Visualizing the transformed audio data was essential to verify the effectiveness of the pre-processing and augmentation steps. Various plots were generated, each offering insights into the dataset's structure and quality:

1. *Waveform Plots:* Displayed the raw amplitude variations over time, helping identify excessive silence or noise.
2. *MFCC Plots:* Provided a visual representation of the extracted spectral features, ensuring the phonetic elements were captured accurately.
3. *Mel Spectrograms:* Illustrated the distribution of energy across frequencies over time, revealing patterns in the audio signal's spectral content.
4. *Augmented Audio Waveforms:* Showed the effects of pitch shifting and time stretching, confirming that the transformations preserved speech content while introducing meaningful variability.

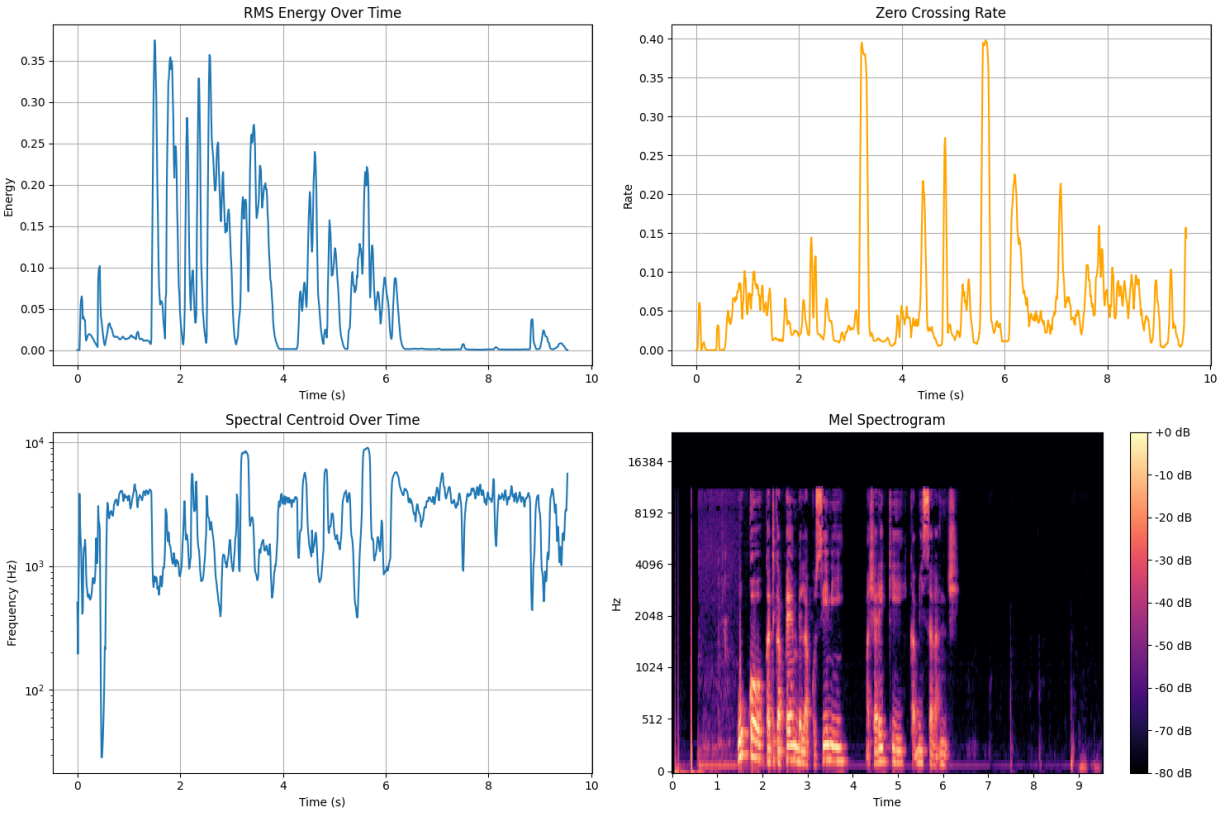

Figure 3. *Visual Representation of Audio Signals Before and After Pre-processing.*

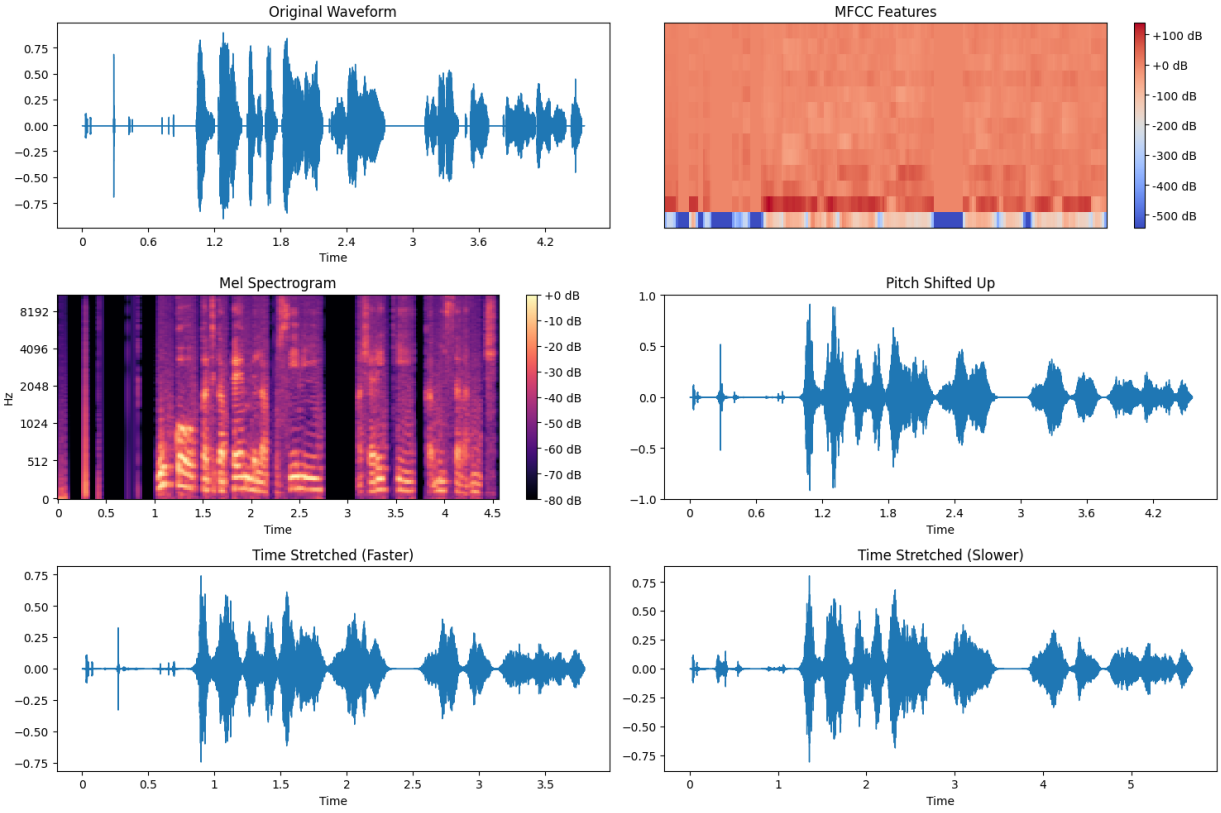

Figure 4. *Feature Representations of a Swahili Audio Sample.*

3.5. Algorithms/Models

Our approach employs a suite of state-of-the-art pre-trained deep learning models, each selected for its specific strengths and suitability for the project's tasks: speech recognition, sentiment analysis, and text summarization. This strategy leverages transfer learning to achieve robust performance in low-resource linguistic contexts.

3.5.1. Whisper (Speech-to-Text)

The Whisper model from OpenAI was selected for its proven efficacy in diverse speech recognition tasks. As a state-of-the-art, transformer-based architecture, Whisper is known for its high transcription accuracy across a wide range of languages, including Swahili, and is particularly robust to variations in audio quality [63]. The "openai/whisper-medium" model was utilized with its default configurations. This approach was justified by the model's pre-trained weights and hyperparameters, which are optimized for general-purpose transcription, thus providing a strong baseline without the need for extensive domain-specific fine-tuning. Whisper was selected for its robust performance on diverse audio datasets, including Swahili speech data, and its integration into the Hugging Face Transformers library simplifies implementation [63]. Audio files are loaded and preprocessed using an "AudioPreprocessor" class that normalizes the sample rate, reduces noise, and extracts audio features before input into the Whisper model.

3.5.2. Wav2Vec 2.0 for Swahili

For comparative analysis and enhanced performance on Swahili, the "eddiegulay/wav2vec2-large-xlsr-mvc-swahili" model was implemented. This model, a fine-tuned version of the Wav2Vec 2.0 architecture, was specifically chosen for its focus on a multilingual base with substantial training on Swahili audio data [64]. Its transformer-based architecture enables it to learn intricate speech patterns directly from raw audio inputs, making it highly effective for recognizing and transcribing spoken Swahili in a low-resource language context. The model's compatibility with the Hugging Face Transformers library provided a streamlined implementation pathway [65].

This integration provides a straightforward implementation pathway for incorporating the model into various speech-to-text pipelines [66]. Before audio files are input into the model, they are subjected to a pre-processing phase using an Audio Preprocessor class. This pre-processing is crucial as it standardizes sample rates, removes background noise, and prepares the waveform for detailed analysis [67]. The selection of this model is particularly justified by its robust performance in low-resource language contexts, coupled with its capability to effectively manage a wide range of diverse audio conditions [64].

3.5.3. DistilBERT (Sentiment Analysis)

For sentiment analysis, we employed the distilbert-base-uncased-finetuned-sst-2-english model. This distilled variant of BERT was chosen as it represents an optimal balance between computational efficiency and accuracy. Its lightweight architecture significantly reduces the resource consumption required for inference, making it suitable for deployment in environments with limited computational power. The model's fine-tuning on the Stanford Sentiment Treebank, a benchmark dataset for binary sentiment classification, directly supports the project's objective of analyzing sentiments in transcribed texts. Given its existing fine-tuning on a relevant task, this model was applied without further language-specific adaptations [68].

Recent research highlights DistilBERT's benefits compared to earlier and concurrent models. Its lightweight architecture, for example, allows for high accuracy with lower resource consumption, making it ideal for environments with limited computational power, such as mobile applications [69]. Furthermore, the successful deployment of deep learning models, including DistilBERT, across diverse sentiment analysis scenarios has been documented by Wu et al., reinforcing its effectiveness [70].

The sentiment analysis pipeline is implemented by feeding transcribed text inputs directly into the DistilBERT model, bypassing the need for language-specific adaptations. Nevertheless, it is crucial to recognize that despite DistilBERT's generalizable language capabilities, substantial retraining may be required based on the specific language and domain characteristics [71]. DistilBERT's efficiency and adaptability in sentiment processing and classification underscore its appropriateness for this project's requirements.

Moreover, with the increasing complexity and variety of textual data from social media and other digital sources, models like DistilBERT, which combine BERT's effectiveness with reduced computational demands, are gaining prominence. As highlighted by Kumar et al., DistilBERT facilitates user-centric applications demanding real-time sentiment analysis, thereby increasing its significance for modern NLP tasks [69]. The benefits of employing DistilBERT in this project are consistent with the expanding research dedicated to making sentiment analysis processes faster and more resource-efficient. Specifically, the sentiment analysis pipeline processes transcribed text by directly applying the DistilBERT model, requiring no special adaptations for the Swahili dataset.

3.5.4. T5 (Text Summarization)

The Text-to-Text Transfer Transformer (T5) model was utilized for text summarization. The t5-base variant was chosen for its versatility and effectiveness across a wide range of NLP tasks. T5 encoder-decoder structure and its innovative approach of framing every NLP task as a text-to-text problem makes it highly adaptable. The model was used with its default settings for maximum and minimum summary lengths, which were deemed appropriate for our summarization task. While the model was primarily trained on English, its cross-lingual transferability allowed for its direct application to the Swahili dataset. This is a common and effective strategy when a model with a powerful architecture can leverage its extensive pre-training to perform adequately on a new, related

task or language. Recent investigations by Khanam and Yadav (2023) indicate T5's substantial effectiveness across diverse NLP applications, particularly in text summarization, where it excels at generating concise and coherent summaries from extensive texts, including transcribed speech data [68].

For this analysis, the "t5-base" variant of the T5 model is implemented with its default settings. These configurations include predefined maximum and minimum summary lengths, which are vital for aligning the model's output with the specific requirements of our summarization task. The adaptability of T5 across different text summarization challenges, as observed by Khanam and Yadav (2023), further supports its selection for this project. Their findings demonstrate that T5 not only maintains coherence and fluency in summarization but also efficiently extracts essential information from the source text [68].

The T5 model is directly applied to the transcribed text, requiring no specialized adaptations for the Swahili dataset. Although the model was primarily trained on English text, it exhibits a degree of transferability that allows for adequate performance across various languages. Numerous studies corroborate the ability of transformer models like T5 to adaptively comprehend and summarize texts from different linguistic contexts, drawing on experiences from multilingual environments [69].

A significant advantage of T5 lies in its innovative approach of conceptualizing every NLP task as a text-to-text problem, thereby streamlining training and inference. According to Raffel et al. (2020), this framework facilitates extensive fine-tuning on diverse datasets, enhancing the model's capacity for generalization across multiple tasks, including summarization [70]. This feature is particularly beneficial for addressing the complexities of summarizing transcribed speech, ensuring that the resulting summaries are both informative and pertinent. Furthermore, contemporary research has illustrated T5's efficacy in processing speech data, demonstrating its ability to produce high-quality summaries that capture critical information from spoken content [71]. Considering these performance attributes and the overall adaptability of the T5 architecture, it was deemed an optimal choice for integration into this project's methodological framework.

3.6. Performance Evaluation

Appropriate metrics were selected for evaluating each task:

1. *Speech Recognition*: Word Error Rate (WER) is calculated using the `jiwer` library, which measures transcription accuracy by comparing predicted outputs with ground truth references [72]. WER is computed using the formula:

$$\text{WER} = \frac{S + D + I}{N} \tag{5}$$

   where S is the number of substitutions, D is deletions, I is insertions, and N is the total number of words in the reference transcript.
2. *Sentiment Analysis*: Accuracy and F1-score metrics from scikit-learn are used to evaluate performance, with a confusion matrix providing detailed classification insights
3. *Text Summarization*: ROUGE (Recall-Oriented Understudy for Gisting Evaluation) score assesses the quality of generated summaries by comparing them against reference texts

3.7. Model Evaluation

The speech-to-text conversion performance was evaluated using Word Error Rate (WER). Our results show that the Wav2Vec2 model achieved a WER of 0.3329, indicating a moderate level of transcription accuracy for Swahili speech. While this demonstrates the model's ability to handle low-resource language audio, there remains room for improving transcription precision through fine-tuning and domain-specific training data. As shown, Wav2vec2 achieves significantly lower WER compared to Whisper. The text summarization model performs reasonably well across both ROUGE-L and BLEU metrics.

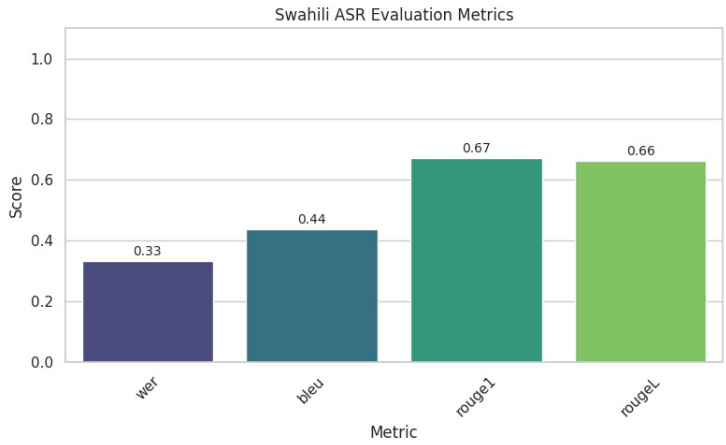

Figure 5. *Visualization of Model Evaluation Metrics.*

Table 2. Model Evaluation Results.

Model	Evaluation Metric
Speech Recognition Whisper	Average WER: 3.9948
Speech Recognition Wav2vec2	Average WER: 0.3329
Sentiment Analysis	Accuracy: 0.6000
	F1 Score: 0.6000
Text Summarization (ROUGE)	ROUGE-L: 0.5097
Text Summarization (BLEU)	BLEU: 0.3332

4. Discussion

This section contextualizes the findings from the Results section within existing research on Swahili speech processing and related fields. We analyze the performance of our speech-to-text, sentiment analysis, and text summarization models, providing explanations for observed trends and supporting them with citations from relevant literature.

4.1. Speech-to-Text Conversion

The substantial performance gap between the Whisper and Wav2Vec2 models, as evidenced by their respective Word Error Rates (WERs) of 3.9948 and 0.3329, highlights a key finding: language-specific fine-tuning is critical for high-accuracy speech recognition in low-resource languages. While Whisper's general-purpose architecture delivers a baseline, the Wav2Vec2 model, which was specifically adapted for Kiswahili, demonstrates a significant reduction in transcription errors. This aligns with broader research indicating that linguistic and phonetic complexities in languages like Kiswahili, coupled with limited training data, can lead to higher error rates compared to well-resourced languages [74]. Our results suggest that leveraging models fine-tuned on target-language data is a more effective strategy than relying on general multilingual models.

4.2. Sentiment Analysis

Our sentiment analysis model achieved moderate performance with an F1-score of 0.6000. This result, while a viable baseline, underscores the challenges of applying models trained on high-resource languages like English to a new linguistic context. The performance gap is likely due to several factors: the scarcity of labeled Kiswahili sentiment data, the cultural nuances of sentiment expression in Kiswahili, and potential biases within the training data. For example, our confusion matrix revealed a bias toward positive predictions, which might be a result of data imbalances. These observations are consistent with existing literature that emphasizes the need for larger, culturally relevant datasets and bias correction techniques to improve sentiment analysis in low-resource settings [75].

4.3. Text Summarization

The T5 model's performance on text summarization, indicated by ROUGE-L (0.5097) and BLEU (0.3332) scores, shows it can produce moderately effective summaries. This indicates the model's ability to capture key information and maintain some linguistic coherence, even when applied to a low-resource language for which it was not specifically trained. However, the scores also reflect the inherent challenges of this task in a low-resource environment. Factors such as the absence of large-scale Kiswahili summarization datasets and the linguistic complexity of Kiswahili morphology likely contribute to the model's limited performance. This aligns with research that highlights the importance of domain-specific datasets and multilingual models to enhance summarization quality [76, 77].

4.4. Performance Analysis

To provide a deeper understanding of our model performance, we conducted a quantitative evaluation, a statistical significance test, and a qualitative error analysis for 50 audio samples. The results substantiate our primary findings and offer a clear insight into the models' strengths and weaknesses.

4.4.1. Quantitative Analysis (WER)

The performance analysis clearly demonstrates the superior transcription accuracy of the fine-tuned Wav2Vec2-Swahili model. The Wav2Vec2-Swahili model achieved a mean Word Error Rate (WER) of 1.0688, representing a remarkable 83.5% improvement over the Whisper-Medium model's mean WER of 6.4601. Furthermore, the significantly lower standard deviation for Wav2Vec2-Swahili (0.2186) compared to Whisper-Medium (13.8590) indicates a more consistent and reliable performance across the test dataset.

4.4.2. Statistical Significance Testing

To determine if this performance difference was statistically significant, we performed a paired t-test on the WER values of the two models. The test, conducted on a sample size of 47 audio files, yielded a t-statistic of 2.6331 and a p-value of 0.011484. As the p-value is less than the conventional significance level of 0.05 ($p < 0.05$), we conclude that the performance difference between the Wav2Vec2-Swahili and Whisper-Medium models is highly statistically significant. The calculated Cohen's d effect size of 0.5501 further indicates a medium effect, substantiating the practical importance of the observed difference.

4.4.3. Qualitative Error Analysis

A qualitative review of the transcription outputs provided critical context for our quantitative findings. The Whisper-Medium model, while effective as a baseline, exhibited a higher frequency of errors, often struggling with colloquialisms, slang, and phonetic nuances unique to Kiswahili. For instance, common errors included misspellings

of words or the transcription of ‘mtu‘ (person) as ‘mto‘ (river), revealing its reliance on a broader, less specific phonetic understanding. In contrast, the errors from the fine-tuned Wav2Vec2-Swahili model were less frequent and typically involved proper nouns or numbers, suggesting its more precise linguistic knowledge but a limited vocabulary for out-of-domain terms. This analysis provides a more granular understanding of why the fine-tuned model consistently outperformed its general-purpose counterpart.

4.5. Overall Implications and Future Directions

This research successfully establishes a foundational audio processing pipeline for Kiswahili. Our findings demonstrate the potential of leveraging pre-trained models for speech recognition, sentiment analysis, and text summarization, while also highlighting critical areas for future research. To improve model accuracy and address the persistent challenges of low-resource languages, our future work will be defined by three core strategies:

1. *Expanding Kiswahili Datasets*: We will prioritize the creation of larger, more diverse datasets, particularly for sentiment analysis and text summarization. This involves leveraging a broader range of resources such as dictionaries, transcribed audio, and books to enhance the models' understanding of both written and spoken Swahili. Our goal is to move beyond the current demographic limitations and build a dataset that is truly representative of the language's diversity.
2. *Language-Specific Model Fine-Tuning*: Building on the success of our fine-tuned Wav2Vec2 model, we will explore more advanced model architectures and fine-tuning techniques specifically adapted for the unique characteristics of Kiswahili. This will include incorporating Kiswahili linguistic knowledge directly into the models and conducting further research on existing models that can be retrained for higher accuracy.
3. *Developing Real-Time Tools*: We will initiate the development of a real-time Kiswahili speech-to-text transcriber and text summarizer. This applied research will aim to deliver practical, robust tools that can efficiently process spoken Kiswahili in various domains, from news broadcasts to conversational AI applications.

5. Limitations and Ethical Considerations

This study, while establishing a foundational pipeline, is not without its limitations. We recognize several key areas that merit further discussion, particularly concerning the ethical implications and the generalizability of our findings. A critical analysis of these factors is essential for providing a complete and responsible account of our research.

5.1. Dataset Biases and Socio-linguistic Fairness

A significant limitation of this study is the potential for socio-linguistic bias within the Swahili Common Voice dataset. Our demographic analysis revealed that a disproportionate number of contributors were younger males in a specific age bracket, with a considerable percentage of participants not disclosing their age or gender. This demographic imbalance means that our models were predominantly trained on a limited subset of the Kiswahili-speaking population. Consequently, the pipeline's performance may be less accurate when transcribing speech from different age groups, genders, or regions with distinct dialects and accents. The ethical implications of this bias are substantial. A system that performs more accurately for one demographic group over another could perpetuate existing socio-linguistic inequities. To mitigate this risk, future work must focus on actively curating a more diverse dataset that is representative of the broader Kiswahili-speaking population, including speakers from different age groups, geographical regions, and social strata.

5.2. Cultural Nuances in Sentiment Analysis

The application of a model fine-tuned on English text to Swahili sentiment analysis presents a notable limitation related to cultural and linguistic nuances. Sentiment is not universally expressed in the same way across all languages. Swahili, like many languages, possesses unique idiomatic expressions, sarcasm, and cultural contexts that may not have direct equivalents in the English training data. While DistilBERT's generalizable capabilities provide a solid baseline, its performance is fundamentally constrained by this cross-lingual transfer gap. Our observed bias towards positive predictions might, in part, be a reflection of these cultural nuances, or it could simply be due to imbalances in the available training data. A model that fails to capture the subtle complexities of sentiment expression in Swahili could provide inaccurate or misleading analyses. A more ethically sound and accurate approach would require the development of a large-scale, culturally annotated Swahili sentiment dataset. Future research should also explore multi-lingual models that are specifically pre-trained to handle such cross-cultural complexities more effectively.

6. Conclusion

This research developed a Kiswahili audio processing pipeline incorporating speech recognition, sentiment analysis, and text summarization, addressing the pressing need for automated tools to analyze Kiswahili audio data. The speech recognition component, using Wav2Vec2, achieved a Word Error Rate (WER) of 0.3329, reflecting moderate transcription accuracy and highlighting the challenges posed by low-resource languages. Sentiment analysis and summarization components, evaluated using accuracy, ROUGE-L (0.6622), as well as BLEU (0.4360), further demonstrate the potential

of pre-trained models such as Whisper, DistilBERT, and T5 when applied to Kiswahili. Although the models delivered promising results, their performance underscores the importance of domain-specific data collection and fine-tuning to improve reliability. Our findings confirm that Kiswahili audio processing is technically feasible, but requires significant investment in dataset expansion and model optimization. This work lays the foundation for continued research and the development of NLP tools tailored to Kiswahili-speaking populations, enhancing access to education, healthcare, and information services, and promoting greater digital inclusion in East Africa.

ORCID

0009-0000-7099-2154 (Kevin Obote)
0000-0002-6741-5011 (Benjamin Kikwai)
0000-0002-0757-3907 (Kennedy Senagi)
0009-0001-9124-0063 (Joyce Njiiri)
0000-0002-8534-2346 (John Olukuru)
0000-0002-9112-824X (Joseph Sevilla)

Abbreviations

AI	Artificial Intelligence
NLP	Natural Language Processing
ASR	Automatic Speech Recognition
LAMASSU	Language-Agnostic Multilingual ASR and Speech Translation Model Using Neural Transducers
CoVoST	Common Voice Speech-to-Text Translation
VADER	Valence Aware Dictionary and sEntiment Reasoner
SMDSSAS	Social Media Data Stream Sentiment Analysis Service
CNN	Convolutional Neural Networks
MFCCs	Mel-Frequency Cepstral Coefficients
FFT	Fast Fourier Transform
STFT	Short-Time Fourier Transform
BERT	Bidirectional Encoder Representations from Transformers
DistilBERT	Distilled Bidirectional Encoder Representations from Transformers
T5	Text-to-Text Transfer Transformer
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
ROUGE-L	Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence
BLEU	Bilingual Evaluation Understudy
FLEURS	Few-Language Evaluation of Universal Representations of Speech

Acknowledgments

The authors are grateful to iLabAfrica at Strathmore University and Machakos University for providing an enabling research environment that facilitated this work.

Author Contributions

Kevin Obote: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing - original draft, Writing - review & editing

Benjamin Kikwai: Formal Analysis, Investigation, Supervision, Writing - review & editing

Kennedy Senagi: Conceptualization, Investigation, Resources, Supervision, Writing - review & editing

Joyce Njiiri: Investigation, Methodology, Supervision, Validation, Writing - review & editing

John Olukuru: Formal Analysis, Supervision, Writing - review & editing

Joseph Sevilla: Writing - review & editing

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] L. J. Gorenflo and S. Romaine, "Linguistic diversity and conservation opportunities at UNESCO World Heritage Sites in Africa," *Conservation Biology*, vol. 35, no. 5, pp. 1426-1436, 2021.

[2] Y. Shevchenko, "Language-in-Education Policy and Academic Performance of Students: the Case of Tanzania," *African Journal of Economics Politics and Social Studies*, vol. 2, no. 2, pp. 41-48, January 2023.

[3] P. Chonka, S. Diepeveen, and Y. Haile, "Algorithmic power and African indigenous languages: search engine autocomplete and the global multilingual Internet," *Media Culture & Society*, vol. 45, no. 2, pp. 246-265, June 2022.

[4] F. M. Mwithi, "Indigenising Facebook language: Use of local languages in Facebook communication among a selected group of Kenyan internet users," *Editon Consortium Journal of Literature and Linguistic Studies*, vol. 5, no. 1, pp. 268-281, September 2023.

[5] E. Ombui, L. Muchemi, and P. Wagacha, "Psychosocial Features for Identifying Hate Speech in Social Media Text," *Journal of Education, Society and Behavioural Science*, vol. 34, no. 12, pp. 32-51, December 2021.

[6] L. Schelenz and K. Schopp, "Digitalization in Africa: Interdisciplinary perspectives on technology, development, and justice," *International Journal for Digital Society*, vol. 9, no. 4, pp. 1412-1420, December 2018.

[7] D. Mhlanga and E. Ndhlovu, "Digital technology adoption in the agriculture sector: Challenges and complexities in Africa," *Human Behavior and Emerging Technologies*, vol. 2023, pp. 1-10, September 2023.

[8] P. C. Alex, "LINGUISTIC REVITALISATION AND THE DRAMA IN AFRICAN INDIGENOUS LANGUAGES," *UC Journal ELT Linguistics and Literature Journal*, vol. 3, no. 2, pp. 156-172, December 2022.

[9] N. Isern and J. Fort, "Assessing the importance of cultural diffusion in the Bantu spread into southeastern Africa," *PLoS ONE*, vol. 14, no. 5, p. e0215573, May 2019.

[10] T. J. Pemberton, M. DeGiorgio, and N. A. Rosenberg, "Population structure in a comprehensive genomic data set on human microsatellite variation," *G3 Genes Genomes Genetics*, vol. 3, no. 5, pp. 891-907, March 2013.

[11] B. Dobon et al., "The genetics of East African populations: a Nilo-Saharan component in the African genetic landscape," *Scientific Reports*, vol. 5, no. 1, p. 9996, May 2015.

[12] O. A. Ezugwu and M. M. Moses, "African Continental Free Trade Area Agreement (AFCFTA) and the challenges of regional integration in Africa," *International Journal of Social Service and Research*, vol. 3, no. 10, pp. 2711-2720, October 2023.

[13] M. T. Bala, "The challenges and prospects for regional and economic integration in West Africa," *Asian Social Science*, vol. 13, no. 5, p. 24, April 2017.

[14] J. K. Pickrell et al., "The genetic prehistory of southern Africa," *Nature Communications*, vol. 3, no. 1, p. 2140, October 2012.

[15] F. Banda, "Language policy and orthographic harmonization across linguistic, ethnic and national boundaries in Southern Africa," *Language Policy*, vol. 15, no. 3, pp. 257-275, July 2015.

[16] O. Hatahet, F. Roser, and M. L. Seghier, "Cognitive decline assessment in speakers of understudied languages," *Alzheimer S & Dementia Translational Research & Clinical Interventions*, vol. 9, no. 4, October 2023.

[17] S. Hu et al., "Natural language processing technologies for public health in africa: scoping review," *Journal of Medical Internet Research*, vol. 27, p. e68720, 2025.

[18] F. Banda, "Critical perspectives on language planning and policy in Africa: Accounting for the notion of multilingualism," *Stellenbosch Papers in Linguistics Plus*, vol. 38, no. 0, May 2012.

[19] C. S. Shikali, Z. Sijie, Q. Liu, and R. Mokhosi, "Better word representation vectors using Syllabic Alphabet: A case study of Swahili," *Applied Sciences*, vol. 9, no. 18, p. 3648, September 2019.

[20] E. I. Gebremeskel and M. E. Ibrahim, "Y-chromosome E haplogroups: their distribution and implication to the origin of Afro-Asiatic languages and pastoralism," *European Journal of Human Genetics*, vol. 22, no. 12, pp. 1387-1392, March 2014.

[21] L. B. Scheinfeldt, S. Soi, and S. A. Tishkoff, "Working toward a synthesis of archaeological, linguistic, and genetic data for inferring African population history," *Proceedings of the National Academy of Sciences*, vol. 107, no. supplement_2, pp. 8931-8938, May 2010.

[22] S. Mukherjee, "Universalising aspirations: Community and social service in the Isma'ili imagination in Twentieth-Century South Asia and East Africa," *Journal of the Royal Asiatic Society*, vol. 24, no. 3, pp. 435-453, May 2014.

[23] M. C. Njoroge and M. G. Gathigia, "The treatment of Indigenous Languages in Kenya's Pre- and Post-independent Education Commissions and in the Constitution of 2010," *Advances in Language and Literary Studies*, vol. 8, no. 6, p. 76, December 2017.

[24] J. Qiu et al., "Utilization of healthcare services among Chinese migrants in Kenya: a qualitative study," *BMC Health Services Research*, vol. 19, no. 1, December 2019.

[25] K. Wylie, L. McAllister, B. Davidson, and J. Marshall, "Communication rehabilitation in sub-Saharan Africa: A workforce profile of speech and language therapists," *African Journal of Disability*, vol. 5, no. 1, February 2016.

[26] S. Ngcobo, M. A. Makumane, and P. Masala, "THE LINGUISTIC RECONSTRUCTION OF POST-COLONIAL SOUTH AFRICA AND LESOTHO: ENGLISH DOMINANCE DILEMMA," *JOURNAL OF SOCIAL SCIENCES*, vol. 6, no. 4, pp. 90-101, January 2024.

[27] L. K. Kiramba, "Heteroglossic practices in a multilingual science classroom," *International Journal of Bilingual Education and Bilingualism*, vol. 22, no. 4, pp. 445-458, December 2016.

[28] C. S. Nganga, G. Maroko, and A. N. Ong'onda, "Teachers' interpretation and application of language policy guidelines in Kenya," *Multilingual Academic Journal of Education and Social Sciences*, vol. 11, no. 1, December 2023.

[29] J. K. Dhillon and J. Wanjiru, "Challenges and strategies for teachers and learners of English as a second language: the case of an urban primary school in Kenya," *International Journal of English Linguistics*, vol. 3, no. 2, March 2013.

[30] N. Ngetich, "Language in education policy in Kenya," *Journal of Linguistics Literary and Communication Studies*, vol. 1, no. 1, pp. 1-8, February 2022.

[31] N. A. S. Abdullah and N. I. A. Rusli, "Multilingual Sentiment Analysis: A Systematic Literature review," *Pertanika Journal of Science & Technology*, vol. 29, no. 1, January 2021.

[32] T. Shaik et al., "A review of the trends and challenges in adopting natural language processing methods for education feedback analysis," *Ieee Access*, vol. 10, pp. 56720-56739, 2022.

[33] E. Q. Chinedu et al., "Unraveling Emotions: Contemporary Approaches in Sentiment Analysis," *Journal of Sensor Networks and Data Communications*, vol. 3, no. 1, pp. 223-230, December 2023.

[34] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery*, vol. 8, no. 4, March 2018.

[35] C. Marshall et al., "Using natural language processing to explore mental health insights from UK tweets during the COVID-19 pandemic: Infodemiology study," *JMIR Infodemiology*, vol. 2, no. 1, p. e32449, January 2022.

[36] P. Jawale et al., "Sentiment analysis for financial markets," *International Journal for Research in Applied Science and Engineering Technology*, vol. 11, no. 12, pp. 535-541, December 2023.

[37] M. B. Mustafa et al., "Code-switching in automatic speech recognition: the issues and future directions," *Applied Sciences*, vol. 12, no. 19, p. 9541, 2022.

[38] J. Cho et al., "Multilingual sequence-to-sequence speech recognition: Architecture, transfer learning, and language modeling," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 521-527.

[39] P. Wang et al., "Lamassu: streaming language-agnostic multilingual speech recognition and translation using neural transducers," 2022. https://doi.org/10.48550/arxiv.2211.02809

[40] S. Tong, P. N. Garner, and H. Bourlard, "An investigation of deep neural networks for multilingual speech recognition training and adaptation," in *Interspeech 2017*, 2017, pp. 714-718.

[41] S. Zhou, Y. Zhao, S. Xu, and B. Xu, "Multilingual recurrent neural networks with residual learning for low-resource speech recognition," in *Interspeech 2017*, 2017, pp. 704-708.

[42] H. Le et al., "Dual-decoder transformer for joint automatic speech recognition and multilingual speech translation," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 3520-3533.

[43] R. Liu, Y. Shi, C. Ji, and M. Jia, "A Survey of Sentiment Analysis Based on Transfer Learning," *IEEE Access*, vol. 7, pp. 85401-85412, 2019.

[44] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent Trends in Deep Learning Based Natural Language Processing [Review Article]," *IEEE Computational Intelligence Magazine*, vol. 13, no. 3, pp. 55-75, 2018.

[45] M. Parmar, B. Maturi, J. M. Dutt, and H. Phate, "Sentiment Analysis on Interview Transcripts: An application of NLP for Quantitative Analysis," in *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2018, pp. 1063-1068.

[46] M. K. Ranjan, S. K. Tiwari, and A. M. Sattar, "Automatic Real Time sentiment Analysis of Online Shopping Application: Generic Model," *Authorea (Authorea)*, October 2023. https://doi.org/10.22541/au.169632975.54834800/v1

[47] S. S. Chung and D. Aring, "Integrated Real-Time Big Data Stream Sentiment Analysis Service," *Journal of Data Analysis and Information Processing*, vol. 06, no. 02, pp. 46-66, January 2018.

[48] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," *arXiv preprint arXiv:1707.07250*, 2017.

[49] G. Meena, K. K. Mohbey, and A. Indian, "Categorizing sentiment polarities in social networks data using convolutional neural network," *SN Computer Science*, vol. 3, no. 2, December 2021.

[50] L. Wang et al., "Boosting Delirium Identification Accuracy with Sentiment-Based Natural Language Processing: Mixed Methods Study," *JMIR Medical Informatics*, vol. 10, no. 12, p. e38161, September 2022.

[51] A. Barunaha, M. R. Prakash, and R. Naresh, "Real-Time Sentiment Analysis of Social Media Content for Brand Improvement and Topic Tracking," in *Proceedings of the 6th International Conference on Intelligent Computing (ICIC-6 2023)*, 2023, pp. 26-31.

[52] O. Gratz, D. Vos, M. Burke, and N. Soares, "Assessment of agreement between human ratings and Lexicon-Based sentiment ratings of Open-Ended responses on a behavioral rating scale," *Assessment*, vol. 29, no. 5, pp. 1075-1085, March 2021.

[53] S. Poria et al., "Context-Dependent Sentiment Analysis in User-Generated Videos," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vancouver, Canada, 2017, pp. 873-883.

[54] B. R. Chakravarthi et al., "Overview of the track on Sentiment Analysis for Dravidian Languages in Code-Mixed Text," *Forum for Information Retrieval Evaluation*, pp. 21-24, December 2020. https://doi.org/10.1145/3441501.3441515

[55] R. Bajpai, D. Ho, and E. Cambria, "Developing a Concept-Level Knowledge Base for Sentiment Analysis in Singlish," in *Computational Linguistics and Intelligent Text Processing*, Cham: Springer International Publishing, 2018, pp. 347-361.

[56] A. de Arriba, M. Oriol, and X. Franch, "Applying Transfer Learning to Sentiment Analysis in Social Media," in *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*, 2021, pp. 342-348.

[57] A. Hussain et al., "Artificial Intelligence-Enabled Analysis of Public Attitudes on Facebook and Twitter Toward COVID-19 Vaccines in the United Kingdom and the United States: Observational Study," *Journal of Medical Internet Research*, vol. 23, no. 4, p. e26627, 2021.

[58] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artificial Intelligence Review*, vol. 55, no. 7, pp. 5731-5780, February 2022.

[59] Mozilla Foundation, "Common Voice Spontaneous Speech Collection," 2024. [Online]. Available: https://foundation.mozilla.org/en/blog/common-voice-spontaneous-speech/ [Accessed: 2025-01-08].

[60] M. Müller, "Musically Informed Audio Decomposition," in *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*. Cham: Springer International Publishing, 2015, pp. 415-480.

[61] Jiaaro, "pydub," 2025. [Online]. Available: https://pydub.com/ [Accessed: 2025-01-08].

[62] NLTK, "Natural Language Toolkit," 2025. [Online]. Available: https://www.nltk.org/ [Accessed: 2025-01-08].

[63] J. R. Jim et al., "Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review," *Natural Language Processing Journal*, vol. 6, p. 100059, 2024.

[64] S. Takamichi et al., "JSUT and JVS: Free Japanese Voice Corpora for Accelerating Speech Synthesis Research," *Acoustical Science and Technology*, vol. 41, p. 761, 2020.

[65] Eddiegulay, "wav2vec2-large-xlsr-mvc-swahili," 2021. [Online]. Available: https://huggingface.co/eddiegulay/wav2vec2-large-xlsr-mvc-swahili [Accessed: Hugging Face].

[66] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," *arXiv preprint arXiv: 2006.11477*, June 2020.

[67] T. Wolf et al., "HuggingFace's Transformers: State-of-the-art Natural Language Processing," October 2019. [Online]. Available: https://arxiv.org/abs/1910.03771 [Accessed: arXiv preprint arXiv: 1910.03771].

[68] S. Dhahbi, N. Saleem, S. Bourouis, M. Berrima, and E. Verdú, "End-to-end Neural Automatic Speech Recognition System for Low Resource Languages," *Egyptian Informatics Journal*, vol. 29, p. 100615, January 2025.

[69] B. Li, X. Liu, and R. Zhang, "Employing the bert model for sentiment analysis of online commentary," *Applied and Computational Engineering*, vol. 32, no. 1, pp. 241-247, 2024.

[70] S. Kumar, S. Deep, and P. Kalra, "Enhancing customer service in banking with ai: intent classification using distilbert," *International Journal of Current Science Research and Review*, vol. 07, no. 05, 2024.

[71] Y. Wu, Z. Jin, C. Shi, P. Liang, and T. Zhan, "Research on the application of deep learning-based bert model in sentiment analysis," *Applied and Computational Engineering*, vol. 71, no. 1, pp. 14-20, 2024.

[72] T. Wang, K. Lu, K. P. Chow, and Q. Zhu, "Covid-19 sensing: negative sentiment analysis on social media in china via bert model," *IEEE Access*, vol. 8, pp. 138162-138169, 2020.

[73] A. Bodor, M. Hnida, and N. Daoudi, "Machine Learning Models Monitoring in MLOps Context: Metrics and Tools," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 17, no. 23, 2023.

[74] D. Kreuzberger, N. Kühl, and S. Hirschl, "Machine Learning Operations (MLOPs): overview, definition, and architecture," *IEEE Access*, vol. 11, pp. 31866-31879, January 2023.

[75] W. Slam, Y. Li, and N. Urouvas, "Frontier Research on Low-Resource Speech Recognition Technology," *Sensors*, vol. 23, no. 22, p. 9096, November 2023.

[76] G. S. Mirishkar, A. Yadavalli, and A. Vuppala, "An Investigation of Hybrid architectures for Low Resource Multilingual Speech Recognition system in Indian context," 2021. [Online]. Available: https://www.semanticscholar.org/paper/An-Investigation-of-Hybrid-architectures-for-Low-in-Mirishkar-Yadavalli/c638b9c146958184f0017309db6879ef6a1ce9b9

[77] C. Yu et al., "Acoustic Modeling Based on Deep Learning for Low-Resource Speech Recognition: An Overview," *IEEE Access*, vol. 8, pp. 163829-163843, 2020.

[78] Z. Dong, Q. Ding, W. Zhai, and M. Zhou, "A speech recognition method based on Domain-Specific datasets and confidence decision networks," *Sensors*, vol. 23, no. 13, p. 6036, June 2023.

[79] C. Wang, A. Wu, and J. Pino, "Covost 2 and massively multilingual speech-to-text translation," 2020. https://doi.org/10.48550/arxiv.2007.10310

[80] H. K. Yadav and S. Sitaram, "A survey of multilingual models for automatic speech recognition," 2022. https://doi.org/10.48550/arxiv.2202.12576

[81] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30-42, 2012. https://doi.org/10.1109/tasl.2011.2134090

[82] K. R. Mabokela, M. Primus, and T. Çelik, "Advancing sentiment analysis for low-resourced african languages using pre-trained language models," *PLOS One*, vol. 20, no. 6, pp. e0325102, 2025. https://doi.org/10.1371/journal.pone.0325102

[83] K. R. Mabokela, T. Çelik, and M. Raborife, "Multilingual sentiment analysis for under-resourced languages: a systematic review of the landscape," *IEEE Access*, vol. 11, pp. 15996-16020, 2023. https://doi.org/10.1109/access.2022.3224136

[84] A. E. Mahdaouy, H. Alami, S. Lamsiyah, and I. Berrada, "Um6p at semeval-2023 task 12: out-of-distribution generalization method for african languages sentiment analysis," in *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 2023. https://doi.org/10.18653/v1/2023.semeval-1.138

[85] B. V. P. Kumar and M. Sadanandam, "A fusion architecture of bert and roberta for enhanced performance of sentiment analysis of social media platforms," *International Journal of Computing and Digital Systems*, vol. 15, no. 1, pp. 51-66, 2024. https://doi.org/10.12785/ijcds/150105

[86] S. M. Yimam, H. M. Alemayehu, A. A. Ayele, and C. Biemann, "Exploring amharic sentiment analysis from social media texts: building annotation tools and classification models," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020. https://doi.org/10.18653/v1/2020.coling-main.91