UNIVERSITY OF NAIROBI

FACULTY OF SCIENCE AND TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE

SWAHILI CONVERSATIONAL AI VOICEBOT FOR CUSTOMER SUPPORT

BY

AWINO EBBIE DORCAS

P52/38818/2020

SUPERVISED BY

DR. ENG. LAWRENCE MUCHEMI

A RESEARCH PROJECT SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS OF THE AWARD OF THE DEGREE OF MASTER OF SCIENCE IN COMPUTATIONAL INTELLIGENCE, DEPARTMENT OF COMPUTER SCIENCE, UNIVERSTY OF NAIROBI

SUBMITTED: 05th JULY 2022

DECLARATION

I declare that this project is my original work done under the guidance and supervision of Dr. Eng. Lawrence Muchemi. This work has not been elsewhere for examination or any other purpose. Instances of use of other people's work within this project have all been properly acknowledged and referenced in accordance with the requirements of the University of Nairobi.

Signed: .... Date: 1ST SEPTEMBER 2022......

Awino Ebbie Dorcas

P52/38818/2020

Declaration by Supervisor

This project has been submitted to the Department of Computer Science, University of Nairobi for examination. This has been done by my approval as the supervisor for this research.

Signed: .... Date: 1) Sep /2022....................

Dr. Eng. Lawrence Muchemi

Department of Computer Science

DEDICATION

This work is dedicated to my father Festus O. Okello for encouraging and supporting me to take on this course and completing it in the shortest time possible.

ACKOWLEDGEMENT

First and foremost I would like to thank the Lord Almighty for the good health, strength, peace of mind and financial breakthroughs throughout this journey.

I am immensely grateful to my research supervisor Dr. Eng. Lawrence Muchemi for his patience, valuable guidance, timely availability, and constant encouragement that has since made the completion of this project a success. Additionally, I am grateful for the mentorship I received from Dr. Edward Ombui, Dr. Lillian D.A.Wanzare and Barrack Wanjawa that improved my knowledge and understanding of my research topic. I would also like to thank my two families for their constant support, prayers, and encouragement throughout this period I have been undertaking my studies. Additionally, I would like to thank my classmates and most especially Martin Ocharo, Geoffrey Sagwe and Velma Namwela for their intellectual support and encouragement when things felt tough. Lastly, I would like to appreciate the entire department of computer science team at the University of Nairobi for their selfless efforts in ensuring that we complete our studies in due time and assisting in our professional growth.

ABSTRACT

Many businesses are now embracing self-service customer support such as conversational AI chatbots. However, most chatbots are internet dependent as they are embedded on websites or internet-dependent apps. Additionally, the available chatbots exploit only major languages such as English rarely used in various rural settings in Africa. This situation can be blamed on the clear disparity in the amount of NLP data for under-resourced languages critical in the development of NLP and AI applications. This research primarily purposed to develop a Swahili conversational AI voicebot for customer support contributing to the survival and development of the cultural and linguistic heritage of Swahili consequently reducing the need for users to learn a new language when interacting with customer support software. The developed voicebot evidenced a WER of 14.56%. During the voicebot's creation, the study collected and analysed NLP data (1000 pattern and response pairs and 3hours of domain-specific speech data) for the Swahili language hence reducing the current gap in the availability of Swahili data for speech and text NLP tasks. The study also identified that while voicebots can be effectively modelled using 3rd party off the shelf solutions and no-code solutions, those targeting under-resourced languages better thrive using customized solutions. This resultant voicebot model will inform and guide the design of similar such voicebots for other domains such as healthcare or insurance.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ADEM	– Automatic Dialogue Evaluation Model
AI	– Artificial Intelligence
ALFFA	– African Languages in the Field: speech Fundamentals and Automation
ASR	– Automatic Speech Recognition
API	– Application Programming Interface
BERT	– Bidirectional Encoder Representations from Transformers
NLG	– Natural Language Generation
NLP	– Natural Language Processing
NLU	– Natural Language Understanding
PC	– Personal Computer
STT	– Speech to Text
TTS	– Text to Speech
UX	– User Experience

1.0 INTRODUCTION

1.1 Background

Businesses often put customer experience at the forefront of their activities primarily due to its close correlation with customer loyalty. Customer experience can be conceptualized as the impression customers have of the business throughout the different levels of the purchasing process (Waqas, Hamzah, and Salleh, 2020). Businesses that get customer experience right at the beginning of the venture often retain a majority of their customers and have these clients recommend the products and services to friends and family. Several studies including that of Ibzan, Balarabe, and Jakada (2016) have evidenced a strong relationship between positive customer experience and willingness to repurchase from the same business. Ahmadinejad (2019) also found that positive customer experience was directly associated with increased word of mouth marketing. Therefore, businesses willing to invest in realizing positive customer experience are deemed to reap immense benefits especially in terms of reputation, revenue, brand advocacy, and loyalty retention.

Customer experience is a sum of several activities that revolve around the product/service and the human resource. While one customer may be blown away by the performance of a product and deem their customer experience as excellent, another will be delighted by the attention they received from the customer support when a problem arose while using the product. This is one basic example that highlights customer experience for any business. Today customers have the power to make or break a business due to the plethora of product variety at their fingertips plus the significant resources such as the internet to ensure they survey for the best business to promote their purchase. Such contributes to the need for businesses to provide remarkable customer experience to ensure customers desire to continue purchasing from them.

Technological advancements have seen businesses embrace several customer support channels to improve their customer service and consequently their customer experience. These avenues include live chats, phone calls, email, and social media. While these channels are all beneficial in their own capacities, the needs and size of the business’s target customers will often make some of the methods unhelpful. Most Kenyan businesses will often target large customer groups contributing to support teams having large ticket backlogs and overwhelming calls especially on Monday mornings as most lack 24/7 customer support.

With the customer support inbox and incoming calls all full, businesses lack better ways of quickly resolving all customer issues resulting in customers waiting for long hours consequently contributing to aggravated anger. Due to such, many businesses are now embracing self-service alternatives through conversational AI, for instance, Chatbots. Conversational AI can reduce these extended waiting time to zero through their instant availability without dependence on human agent availability. Customers get their solutions without having to remain on hold as they wait for human agents or being switched between different people for solutions. Other than the quick support, conversational AI ensures customers are attended to round the clock thus allowing businesses to remain available all day long despite being physically closed.

1.2 Problem Statement

While businesses have in the recent past embraced chatbots for their customer support, most are internet dependent as they are embedded on websites or internet-dependent apps. Such excludes users in remote areas from accessing these conversational AI agents due to poor internet connectivity and lack of highly technological devices such as smartphones, laptops, or personal computers (PCs). Additionally, these chatbots exploit only major languages such as English that are rarely used in several rural settings in Africa. Such, therefore, poses a challenge for that rural setting resident who manages to acquire a smartphone but is illiterate in the English language.

However, the challenge persists as this situation becomes difficult to resolve as modern AI technologies are only capable of producing highly complex spoken and written languages for high-resourced languages such as English and Mandarin. Such is as a result of the clear disparity in the amount of NLP data for under-resourced languages critical in the development of NLP and AI applications for these languages. Developing NLP applications is, therefore, challenging for resource-poor languages such as Swahili due to limited resources. Low-resourced languages are in dire need for resources to overcome these barriers to advance NLP tasks within these languages.

1.3 Objectives

Overall Objective

This study’s overall objective is to develop a Swahili voicebot capable of performing customer support tasks.

Specific Objective

1. To collect and analyse NLP data for the Swahili language
2. To develop a model for Swahili voicebots
3. To test the validity of the proposed model

1.4 Research questions

1. What are the best approaches to the modelling of a Swahili voicebot?
2. What are the best methods to test the validity the proposed model?

1.5 Significance of the study

Most businesses currently employ English-based conversational AI thus Swahili voicebots hold the potential to provide businesses a wider customer reach as the voicebots will be non-internet dependent and in a language understood by more customers unlike English. Also this study will contribute to the survival and development of cultural and linguistic heritage of Swahili by developing an application that uses Swahili reducing the need for users to learn a new language and forget their native language. Additionally, this study is significant as it reduces the current gap in the availability of Swahili data. Such will see future researchers have free and accessible data at their disposal that they can use to solve various real-life AI and machine learning problems using the Swahili language.

2.0 LITERATURE REVIEW

2.1 Factors Hindering Adoption of Conversational AI

Radhakrishnan and Chattopadhyay (2020) in their study identified cost of adoption, performance, lack of prior experience with AI, and lack of infrastructure as some of the barriers to adoption of AI. Lack of infrastructure especially absence of stable network connectivity is a critical hindrance to the adoption of AI in Africa and Kenya in particular. Reddick et al. (2020) argue that while broadband access continues in first world countries such as the United States, several third world countries continue to lack this access that has become even more apparent with the coronavirus pandemic. With the sustainable development goals (SDGs) prioritizing digital inclusion as one of its critical goals, Reddick et al. (2020) asserts that countries are currently struggling to close the existing digital divide to ensure their citizens exploit online services. Conversational AI like other frontier technologies heavily relies on steady and reliable internet connections in a society where half of Kenya's population remains offline. Kenya currently lacks adequate digital infrastructure and for most of its population internet costs are prohibitive to the use of conversational AI services.

Previous research primarily focused on web-based chatbots accessible only via smart devices such as laptops, PCs, or smartphones. Existing chatbot technologies in Kenya today include Safaricom's Zuri, UBA Kenya's Leo, and Jubilee Insurance's Julie that perform several customer support tasks and digital sales operations. Digital divide also persists in terms of the ownership of technological devices capable of accessing the internet. For a majority of rural residents smart devices remain a luxury only afforded by a few. Such restricts these users to simple phones dubbed *kabambe.* These phones only offer the basic mobile phone services including sending and receiving calls and text messages and will rarely offer reliable access to web browsers. Nyongesa, Omieno, and Otanga, (2020) contend that the slow adoption of AI in Kenya primarily stems from the limited knowledge and experience with AI and its significance in expanding business operations. The authors also argue that consumer's today tend to overvalue their experience rather than a business's products/services.

2.2 NLP for African Languages

Karakanta, Dehdari, and van Genabith (2018) argue that NLP plays a significant role in the persistence of human language through the development of tools and applications that different speakers use in their native languages. There is significant need for people to use their own language to socialize, learn, and interact with each other without being forced to learn new languages. Most frontier technologies are developed using high-resourced languages such as English due to availability of large unlabelled and labelled datasets that lack for low-resourced languages like African Languages. Magueresse, Carles, and Heetderks (2020) identify that NLP research has since exploited only 20 languages of the available 7000 languages in the world. Such situations persist despite a majority of world's population lacking significant literacy in English language which is the most used language in NLP applications. Additionally, these high-resourced languages are often centralized to small regions and hold various similarities that hinder generalization of developed NLP models to low-resourced languages that have an array of variances, for instance, presence of agglutination as witnessed in the Swahili language.

African languages, Swahili in particular, are often disadvantaged in their implementation in machine learning and AI tasks due to data inadequacy within the language for NLP resulting in their classification as low-resourced languages. Shikali and Mokhosi (2020) hold that Swahili, a Bantu language spoken with popularity in East and Central Africa currently has limited pre-processed open access data that consequently hinders NLP research on the language. Researchers are today forced to use Swahili data from the Helsinki corpus developed by the Helsinki University in collaboration with the University of Nairobi. Shikali and Mokhosi (2020) argue that while other corpuses exist such as that provided by Gelas, Besacier, and Pellegrino (2012), these datasets are often unannotated and require significant pre-processing to ensure the development of accurate and efficient NLP models capable of comparing to the available state-of-the-art NLP models for high-resource languages.

2.3 Conversational AI

Artificial Intelligence (AI) includes the mimicking of human cognitive functions including reasoning, social intelligence, constant learning and problem solving. Panesar (2020) asserts that conversational AI intends to provide customers with seamless, simple, and smooth experiences. Conversational AI includes a set of technologies that realize automated speech and messaging applications mimicking human-like interactions between computer programs and human beings. These technologies allow human-like communication through speech and text recognition, intent and different languages comprehension, and offering of responses that imitate actual human-like conversations (Panesar, 2020). Conversational AI solutions utilize both text and voice modalities in the form of chatbots and voicebots evidenced in web chats and smart speakers among others. While conversational AI remains a significant research area, the best conversational AI solutions target offering seamless communications indistinguishable from conversations delivered by actual humans.

2.3.1 History and Evolution of AI conversational Agents

Conversational AI technologies began in the 1950s after Alan Turing developed ELIZA the first chatbot meant to communicate with people without them realizing they were conversing with an artificial interlocutor. However, ELIZA could not effectively carry out long conversations, learn from previous conversations to inform future responses, and had a limited knowledge that resulted in its prowess in only few domains (Assenmacher et al., 2020). Despite significant limitations evidenced with ELIZA, the chatbot significantly inspired research on conversational AI. ELIZA utilized pattern matching and a template response scheme to converse with users (Assenmacher et al., 2020).

In 1972, PARRY a chatbot designed to act as a schizophrenic patient was developed. Unlike ELIZA, PARRY had a personality and despite its basing its responses from pre-defined assumptions, it also deciphered pre-defined emotional cues based on weights within a user's utterances that improved its responses (Adamopoulou and Moussiades, 2020). All in all, PARRY had low language capabilities, was slow and failed to learn from previous conversations.

Further, in 1995 ALICE, the first online chatbot inspired by ELIZA was developed. Unlike ELIZA, ALICE could handle longer discourses and covering any domain due to its increased knowledge-base (Adamopoulou and Moussiades, 2020). These differences are attributed to the use of Artificial Intelligence Markup Language (AIML) that was a new programming language crafted primarily for ALICE. On the downside, ALICE could not generate human-like conversations that embody emotions and attitudes (Adamopoulou and Moussiades, 2020). In the 2000s technological advancements on chatbots significantly improved resulting in the development of personal assistants embedded into smartphones and various home devices and could understand human voice commands. These developments include Amazon's Alexa, IBMs Watson, Apple's Siri, Microsoft's Cortana, and the Google Assistant.

Chatbots are of various categories including goal-oriented, permission-oriented, service-oriented, knowledge-oriented, and response-oriented among others (Adamopoulou and Moussiades, 2020). Goal oriented chatbots can further be classified under informative chatbots, conversational chatbots, and task-based chatbots. While informative chatbots offer specific information stored in fixed sources, task-based chatbots offer task-related assistance to users (Adamopoulou and Moussiades, 2020). Lastly, conversational chatbots offer natural and human-like conversations with users (Adamopoulou and Moussiades, 2020).

2.3.2 Conversational AI Technologies

Conversational AI utilizes various technologies to listen, comprehend, learn, react, and respond to interactions with human users.

Automatic Speech Recognition (ASR)

While humans communicate in various ways including gestures and facial expressions, speech is considered the commonest mode of communication. Conversational AI applications often receive information either through text or voice. Through ASR these applications make sense of the spoken phrases and consequently translate them to machine readable formats for further processing. Benkerzaz, Elmir, and Dennai (2019) define ASR as the graphical representation of speech frequencies emitted by humans over a function of time. The primary goal of ASR is to allow computer programs listen and understand acoustic data for further information processing.

Natural Language Processing (NLP)

NLP is the fundamental of the recognition of language by conversational AI agents. NLP allows applications to understand text and speech inputs through natural language understanding (NLU) and natural language generation (NLG). NLU reads and processes text to determine its meaning. Nyongesa, Omieno, and Otanga (2020) hold that NLU in text processing is often the most complex of the two NLP tasks as speakers will often use abbreviations or slang that are absent in the language models provided by the ASR realizing challenges in accuracy. Therefore the most effective NLU algorithm is one that caters to these speaker tendencies without interfering with the smooth flow of the conversation. Ayanouz, Abdelhakim, and Benhmed, M. (2020) state that NLU uses five key steps to decipher meaning behind text or speech inputs including lexical analysis, syntactic analysis, semantic analysis, discourse integration, and pragmatic analysis. NLG, on the other hand, involves the formation of linguistically correct texts to realize a comprehensible response.

Speech Synthesis (Text to Speech)

Depending on the conversational AI application's architecture, either a text or voice response is provided. Speech synthesis includes the artificial production of human speech over voice modality (Wu, Zhao, and Zhang, 2020). Speech synthesis allows for the conversion of written text compiled through the NLG module to speech.

Machine Learning

This technology is responsible to the learning and improving of the conversational AI application over time. The application constantly accepts corrections and learns from its past interactions to realize better responses in future conversations.

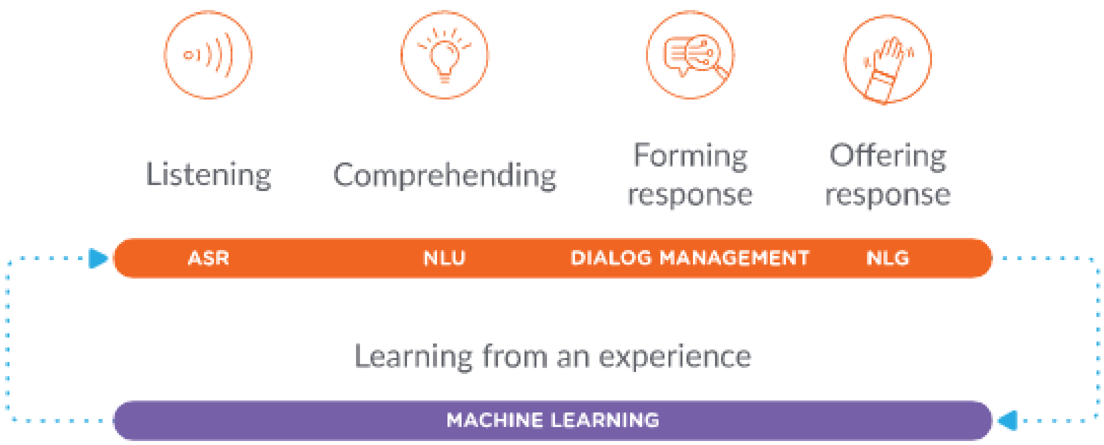

Figure 1: Basic architecture of conversational AI applications (source: https://www.interactions.com/conversational-ai/)

2.3.3 Voicebots

Orungati (2020) describes a voicebot as a virtual bot that conducts conversations through voice. Voicebots are designed to mimic human-like speech and conversational patterns for various purposes including customer support. Orungati (2020) holds that while sophisticated voicebots will employ AI and NLP technologies to understand users and appropriately respond, simpler versions will often just scan through keywords and match them to words within a pre-defined database and provide responses. Voicebots are simple chatbots that use voice as they utilize the underlying chatbot design. Similar to chatbots, voicebots are significantly evolving in various spheres as they mostly target illiterate users who cannot chat with bots in written formats. Unlike chatbots, voicebots offer more flexibility as the user can multitask as they converse with the bots.

2.4 Approaches Used to Develop Conversational AI Applications

Conversational AI are often developed using two distinct approaches: pattern matching and machine learning approaches.

2.4.1 Pattern Matching

ELIZA and ALICE were the pioneer chatbots to utilize pattern matching. Adamopoulou and Moussiades (2020) contend that the pattern matching approach allows the conversational AI agent to match user input to a pre-defined rule within its rule-base and identifies the most appropriate answer from the pre-defined responses. Pattern matching is often performed using three distinct languages including AIML, Chatscript, and Rivescript. Adamopoulou and Moussiades (2020) state that the primary advantage of the pattern matching technique is its high response time as the rule-based systems do not perform deep semantic or syntactic evaluation of the inputs. The pattern matching technique typically does not create new responses to questions as the rules and conversational patterns within the rule-based system are pre-coded by the developer. Additionally, responses from conversational agents using the pattern matching technique lack originality and spontaneity evidenced in real human responses.

2.4.2 Machine Learning

Adamopoulou and Moussiades (2020) state that machine learning is a technique used to develop conversation AI agents that includes the use of NLP to extract vital information from the user that informs responses and allows the system to learn from conversations that influence future responses. Conversation AI agents using machine learning often use artificial neural networks (ANNs) to retrieve previously learned data and provide scores that are then used to pinpoint the most likely response.

2.5 Interaction Styles of Conversational AI Agents

Conversational AI agents are currently adopted in various domains for a multitude of purposes that require different interaction styles. Paikari and Van Der Hoek (2018) identified four distinct interaction styles including the dull style, the alternate vocabulary style, the relationship building style, and the human-like style. The dull interaction style involves the use of single response words or phrases that are regularly repeated. The alternate vocabulary style, on the other hand, involves the use of a larger response variation base, that is, the agent offers the same answer but using synonyms. The relationship building style refers to the agent’s use of vocabularies based on the conversational atmosphere; the bot can easily shift from a professional tone to a more laid-back tone. Additionally, bots employing this style significantly control the conversational flow and can easily throw in humour in the form of jokes. Lastly, the human-like interaction style involves the learning from past conversations to influence future discussions using more subtle and human-like conversational patterns that influence meaningful dialogue.

2.6 Evaluating Conversational AI models

Currently the common approaches in the evaluation of conversational AI include the use of word overlap metrics including BLEU, ROUGE, or METEOR and human judgement. Lowe et al. (2017) identified that unlike other evaluation techniques, human judgement was time consuming and expensive. While BLEU and METEOR are primarily used in machine translation models, ROUGE is applied in the evaluation of text summarization models. However, Lowe et al. (2017) argue that these automatic evaluation metrics are often biased and poorly correlate with human judgements of the quality of conversational AI responses. To mitigate this challenge Lowe et al. (2017) developed ADEM that is an evaluation model that learns the prediction of human-like scores in its evaluation of conversational AI models. Lowe et al. (2017) contend that rapid prototyping and testing of new conversational AI models can be realized through use of accurate automatic evaluation metrics as they consequently decrease the cost of employing human evaluation of the developed models. The authors in their study highlight ADEMs high accuracy that supersedes that of evaluation metrics such as BLEU. Other significant evaluation techniques used in several previous studies include the BERTScore and the BLEURT. Clinciu, Eshghi, and Hastie (2021) established that BERTScore and BLEURT outperformed word-overlap metrics and realized higher correlations with human evaluations with score ranges of 023-0.43 and 0.26-0.53 respectively.

2.6.1 Word Overlap Metrics

Lowe et al. (2017) hold that word overlap metrics fail to effectively capture semantic similarities between the reference and model responses particularly in instances that lack common words. The authors argue that while this drawback is less critical for machine translation and summarization tasks, it is of immense significance in the evaluation of conversational AI models. Unlike in machine translations where one can infer the quality of a translated sentence by assessing the word-overlap between words, in conversational AI models such becomes challenging as one has to consider the context as the most reasonable response will be context-dependent. Lowe et al. (2017) assert that conversational AI models include significantly high response diversity that is often uncaptured by word-overlap comparisons.

2.6.2 An Automatic Dialogue Evaluation Model (ADEM)

ADEM unlike word-overlap metrics captures semantic similarities and utilizes context and reference response to calculate the model's score. Lowe et al. (2017) hold that the metric's hierarchical structure ensures that it learns from previous conversation responses to better understand the context.

2.6.3 BERTScore

BERTScore is an evaluation metric for NLG that calculates cosine similarity scores between tokens in the model sentence and the reference sentence. Devlin et al. (2018) states the BERTScore is developed based on pre-trained BERT contextual embedding. Zhang et al. (2019) attest that unlike word-overlap metrics, BERTScore underscores the significance of matching paraphrases and distance dependencies and word ordering. Additionally, Zhang et al. (2019) contend that the evaluation metric highly correlates with human evaluation of conversational AI models.

2.6.4 BLEURT

Sellam, Das, and Parikh (2020) identified BLEURT as a NLG evaluation metric based on the BERT model that exploits pre-trained synthetic data and utilizes random perturbations of sentences stored under Wikipedia that represent sentences with diverse lexical and semantic level supervision goals.

2.7 Conversational AI with humanness

Autonomous agents including the conversational agents geared towards exemplifying human like behaviour are developed using the BDI architecture. This model includes three distinct concepts: beliefs, desire, and intentions. The designing of conversational AI agents that replicate humanness requires the incorporation of these three concepts. The aspect of belief allows the agent to provide knowledge based on the current conversation's context (the environment) and its internal state of mind. The concept of desire, on the other hand, represents the goals the agent wants to achieve. For example, within the domain of customer support the agent's goal is to solve the customer's issue in the most appropriate manner.

Lastly, the aspect of intention refers to the sequence of steps the agent chooses to follow to achieve its desires.

While the BDI architecture has in the past shown its significance in the development of human-like characters, Svenningsson and Faraon (2019) contend that the model lacks several generic aspects of human-like reasoning and behaviours. Additionally, while the simulation of human-like behaviour is critical for conversational AI agents, Svenningsson and Faraon (2019) argue that the level of humanness may sometimes feel "off" that adversely affecting user-experience (UX). Therefore, if the level of humanness evidenced in the conversational agent does not match conventional and unwritten rules of human interaction then the users may be dissatisfied with its use. Jain et al. (2018) established that users appreciated conversational AI agents that exemplified humanness through comprehension of negative statements, adherence to the conversational context, failure management, asked intelligent questions and understood the significance of turn-taking.

2.8 Related Works

Several scholars have in the recent decade highlighted the value of convenience of conversational AI in various domains and research on this area remains at the forefront to extensively evaluate their efficacy and user experiences. Such suggests the significance of the rise of conversational AI technologies to positively influence growth and seamless user experience in several domains. More specifically, the recent burst in research on conversational agents has led to the substantial increase in research on chatbots, their use, their efficiency, user behaviour, user perceptions, and user preferences.

For example, Nadarzynski et al. (2019) in their study evaluated user acceptability of conversational AI chatbots in the healthcare sector that established that majority of users were hesitant with the adoption of chatbots in healthcare. The respondents cited their non-understanding of how chatbots could accurately respond to user health-related queries and concisely decipher signs and symptoms of various diseases and more specifically those of less common illnesses. Such could significantly increases the chances of harm to patients through misdiagnosis. Additionally, Nadarzynski et al. (2019) established that users previously experienced lack of empathy and incomprehensibility of emotional issues by the chatbots that could adversely impact mental health patients.

On the upside, Nadarzynski et al. (2019) found that users were more motivated to use chatbots unlike medical phone helplines as they were time efficient and they could better comprehend written medical advice. Smutny and Schreiberova (2020), on the other hand, evaluated chatbots in the educational sector. The study established that 46% of educational chatbots lacked proper simulation of human-like dialogue patterns. The study also found that only 49% of these chatbots embraced a conversational tone, attitude, and style similar to human beings.

Nyongesa, Omieno, and Otanga, (2020) in their study identified that while Kenyan businesses had to a small extent embraced AI technologies, the telecommunication industry was in fact ready to exploit the benefits of conversational AI chatbots for customer support. The authors also established that complex technologies associated with conversational AI hindered its use among Kenyan businesses and therefore realized a less complex and highly compatible prototype. Such would then ensure seamless incorporation to already existing telecommunication infrastructure.

In yet another study targeting implementation of chatbots in the German insurance sector, Rodríguez Cardona et al. (2019) established that potential consumers of the chatbot technology and practitioners were still unsure of the ability of chatbots to process complex insurance decisions without considerable human support. Additionally, Rodríguez Cardona et al. (2019) found that insurers were sceptical of the chabot’s ability to mimic the convincing and influential nature of human agents to the point to influencing customer purchasing decisions.

2.9 Research Gap

Recent research and developments on conversational AI have realized applications capable of executing a wide array of complex tasks but with provisions only for major languages that are high-resourced. In Africa, Kenya in particular, consumers of AI applications are particularly conversant with vernacular as it is their first language. Despite such, conversational AI in Kenya persists in their use of English that hinders usage by a majority of the Kenyan population especially those in rural areas. Additionally, there is a significant disparity in the presence of research on the use of Swahili speech evidenced in conversational AI agents.

2.10 Conceptual Design

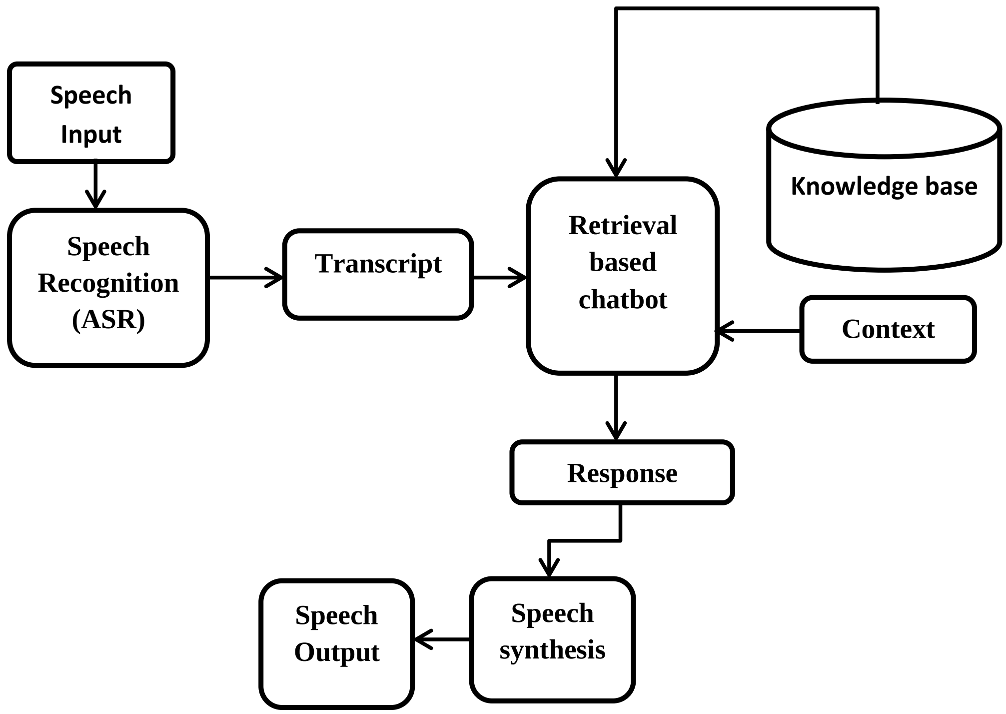

Figure 2: Conceptual Design

3.0 METHODOLOGY

3.1 Introduction

This chapter details the approach that will be taken to conduct this study. It outlines the methodology, data sources the expected contributions.

3.2 Methodology

The core of this system (the chatbot) uses a retrieval based methodology with emphasis on the customer support domain. This closed domain system uses directed graphs where the model is trained to select the best ranked response from a finite set of predefined responses. The query patterns and responses are inputted manually by the developer before training begins. The system is trained in a manner that the model receives a set of queries commonly asked by the user in multiple ways and also a set of corresponding responses.

A retrieval-based methodology allows the developer to control user experience and match it to their expectations of usability and functioning. Additionally, Surendran, Murali, and Babu, (2020) assert the optimality of a retrieval based methodology for conversational AI development for closed domain systems. Therefore, given a user input utterance as the input to the chatbot the system will determine its context based on the primary intent within the input and match it to the candidate responses after evaluation of the features between the input and patterns (possible user queries inputted during training). These feature representations are done using recurrent neural networks (RNNs).

Additionally, the chatbot uses speech inputs which are passed over an ASR for decoding to obtain the textual queries. The ASR is developed using CMU Sphinx which is an open-source ASR toolkit. This toolkit offers customization of the language model, dictionary, and acoustic model that are critical in the decoding process unlike other ASR platforms like Google Speech to Text API or IBM Watson.

Pseudocode

Step 1: *Categorize the FAQ data based on intents*

Step 2: *Place the Questions under the patterns tag*

Step 3: *Place the possible Answers to the questions under the response tag*

Step 4: *Pre-process the data using NLP techniques for example, lemmatization, lowercasing of the words, and removing of duplicates*

Step 5: *Train the chatbot*

Step 6: *Create audio files and divide into two (70% train and 30% test)*

Step 7: *Use the audio files created to create file_ids list, transcript, phone_list, dictionary*

Step 8: *Create language model using SRILM toolkit*

Step 9: *Use CMU Sphinx toolkit and the above created files in steps 1-4 above to train and test the acoustic model*

Step 10: *Use CMU Sphinx toolkit to calculate the WER and SER of the model*

Step 11: *Create a GUI that uses:*

- *the chatbot model created under step 5*
- *language model under step 8*
- *acoustic model created under step 9*
- *Google Text-to-Speech API*

Figure 3: Voicebot Pseudocode

3.3 Data Sources

Chatbot Data

Chatbots require training datasets with a conversational flow, for example, sentences that build on each other's content and context or questions and responses. This project opted for the question and response conversational flow hence prompting a look into various open-source FAQ datasets. While several open-source FAQ datasets exist across the internet, a majority were open-domain datasets making them unsuitable for use within this project. Additionally, some like the Ubuntu Dialogue Corpus includes heavy technological terms limiting effective translation to Swahili for application in this project. Therefore, this project obtained its data from Safaricom's FAQs section on the organization's website. This data source was advantageous due to its hierarchical structure with already curated question and answers pairs and hence prompted less processing. Also, these FAQs were originally in English but were translated to Swahili using Google Translate for proper application into this project. It is a simple dataset with 88 manually created tags denoting various customer intents. Each tag comprises of patterns and responses. The patterns and responses take a question and answer format respectively.

	Attribute	**Description**
1.	Tags	These are keywords that denote the intent within user queries
2.	Patterns	These are sample queries in multiple approaches that a user may ask with regards to a particular tag
3.	Responses	These are the answers generated by the bot with regards to a particular tag
4.	Context	These are scenarios that are provided within the dataset in case a search and find operation is needed

Table 2: Dataset description

The created dataset has a format as follows:

```
{"intents": [
   {"tag": "salamu",
    "patterns": ["Habari","Halo", "siku njema","Waambaje?","vipi"],
    "responses": ["Halo, nikusaidieje?",  "Mzuri, nikusaidieje?"],
    "context": [""]
    },
…
   ]
}
```

Figure 4: Excerpt of the chatbot dataset format

This dataset includes text data with 88 manually created tags and multiple patterns and responses. During manipulation of this data various text pre-processing of the text data occurred before creation of the deep learning chatbot model. Some of the pre-processing activities include tokenization using NLTK and lemmatization. These intents assist the chatbot the user's desired goal, for example, in this project the intents describe actions like getting soft mobile loans, credit purchase, and call flashbacks among others.

ASR Data

An in-house corpus was created from the chatbot text data with one male and two female speakers uttering these words. The creation of an in-house speech corpus arose due to the challenges of proper speech recognition accuracy of customer support terms as the available speech corpuses are large open domain with minimal customer support phrases. CMU Sphinx toolkit used for this speech recognition task is specific on the recording requirements for the audio files. These requirements are as follows:

Parameter	**Value**
Sampling rate	16kHz
Bit rate	16bits
Audio Channel	Monophonic (Mono)
Wave format	.wav

Table 3: Recording system parameters, used for the corpus preparation

3.4 Expected Contributions

This study expects that it will reduce the current gap in the availability of Swahili data for speech and text NLP tasks. This will ensure that future researchers have free and accessible data that they can use to solve various real-life problems, for instance, that related to conversational AI and question answering using the Swahili language. Also this study anticipates contributing to the survival and development of cultural and linguistic heritage of Swahili by developing an application that uses Swahili reducing the need for users to learn a new language and abandoning their native language. Additionally, this study anticipates that the resultant voicebot model will inform and guide the design of similar such voicebots for other domains such as healthcare or insurance. Similarly, this research envisions informing and contributing to future research on conversational AI particularly in the determination of better approaches to the development of conversational AI models.

4.0 ANALYSIS, DESIGN & IMPLEMENTATION

4.1 Introduction

This chapter details the system's requirements, the design and finally the implementation of system using the methodology detailed in chapter 3.

4.2 Architectural System Design

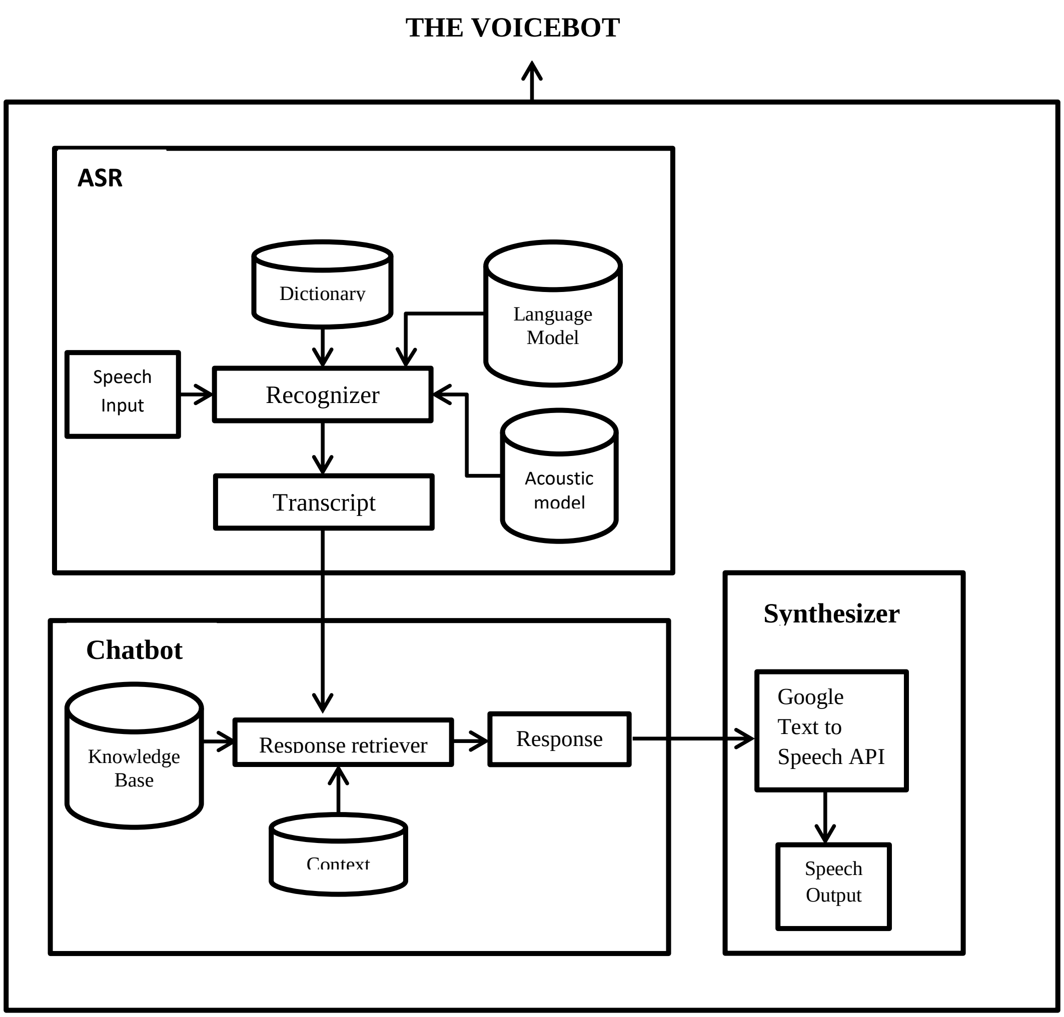

Figure 5: Architectural system design

4.2.1 ASR (Speech to Text)

Swahili is a Bantu language now recognized as a regional language following its widespread use in most countries in the East African region. Swahili comprises of five vowels (a, e, i, o,

u) and 27 consonants (b, ch, d, dh, f, g, gh, h, j, k, kh, l, m, n, ng, ng', ny, p, r, s, sh, t, th, v, w, y, z)

Language Model

The language model is a representation of word usage within a language and is normally customized for each application. High speech recognition accuracies often occur when the language model includes a wider proportion of the words utilized in the acoustic model training. For this project the language model was trained using the chatbot text data and tested using unseen data using the SRI language modelling toolkit (SRILM).

The Dictionary

The training of the acoustic model was also done in the presence of a dictionary. The dictionary is a mapping of each word within the speech corpus to its pronunciation. The following table shows and excerpt of the dictionary used in the training and recognition phases:

Word	**Pronunciation**
unaelewa	UH N AH EH L EH W AH
vyote	VY OH T EH
wateja	W AH T EH J AH
malipo	M AH L IH P OH

Table 4: Excerpt of the dictionary

Acoustic Model

The acoustic model is a statistical representation of the words

An acoustic model represents statistically a range of possible audio representations for the phonemes. During training feature files are created for each of the audio files which include a sequence of feature vectors consisting of the Mel-Frequency Cepstral Coefficients (MFCCs). The training process includes the below steps:

1. Obtaining the speech corpus for training alongside its utterance transcription
2. converting the audio data to a sequence of feature vectors
3. converting the text into a stream of linear HMMs using their pronunciations as per the dictionary

4. determining the best state alignment through the sentence HMM
5. develop a suitable statistical model with the collection of feature vectors. The circularity in this training process is resolved using the iterative Baum Welch or forward-backward training algorithm

The Recognizer (Sphinx-4)

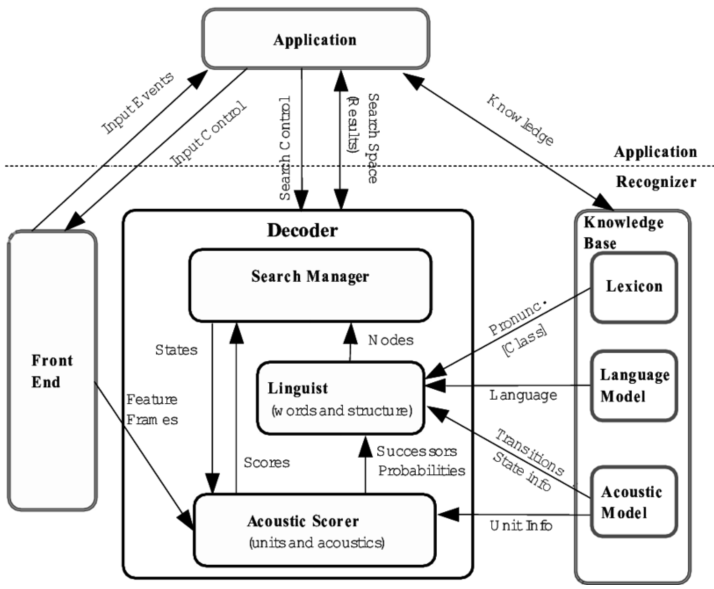

Figure 6: Sphinx-4 Architecture (source: Lamere et al.,2003)

4.2.2 Speech Synthesis (Text to Speech)

This project utilized the Google Cloud Text to Speech API for its speech synthesis as opposed to a custom speech synthesis model. The Google Cloud Speech API allows developers to convert text input into audio output through seamless application of neural network models within the API. The API has a wide voice selection of approximately 220 people across over 40 languages. Despite this wide inclusivity, the API lacks a Swahili speaker prompting this system to the use of an Arabic speaker due to the similarity between Swahili and Arabic.

Figure 7: Google Text to Speech Design (source: Bisong, 2020)

4.3 System Results

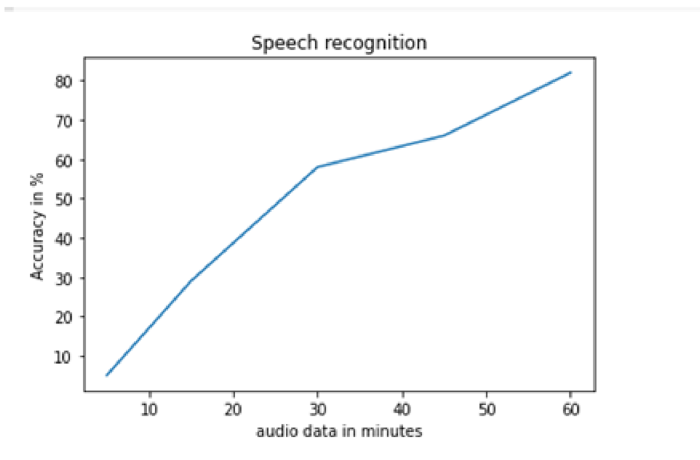

Figure 8: Acoustic Model training Results

After successful compilation of the codes the model is trained using 200 epochs with 96% accuracy.

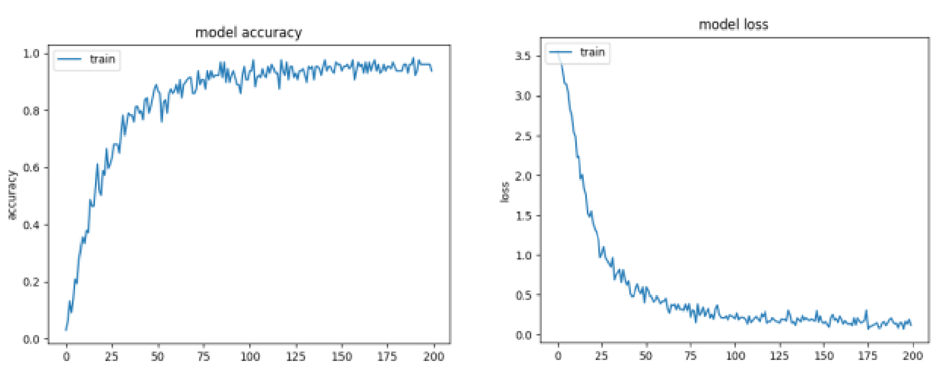

Figure 9: Model accuracy and loss results for the chatbot training

Correct Sentence	Transcribed Sentence
Habari yako?	habari yako
Ningependa kuelewa m-shwari	ni ya mtandao call ya wake m-shwari
Mahitaji ya m-shwari ni zipi?	mahitaji nitajuaje vipi
Je, kuna tozo zozote wakati natumia m-shwari?	je una tozo zozote wa hati za je ni wake
Na je, mtu anaweza kulemaza m-shwari?	na je mtu anaweza ninawezaje wa
Asante kwa kunisaidia	asante call nini je

Table 5: Excerpt of the transcribed sentences

Below is a screenshot of the outputs from the model in terms of text. The voice section would not be reproduced in writing. As evidenced below while the ASR understands the words “asante” and “nimeshukuru”, the chatbot model correctly deciphers them as questions or comments hence prompting the response of “tafadhali rudia”

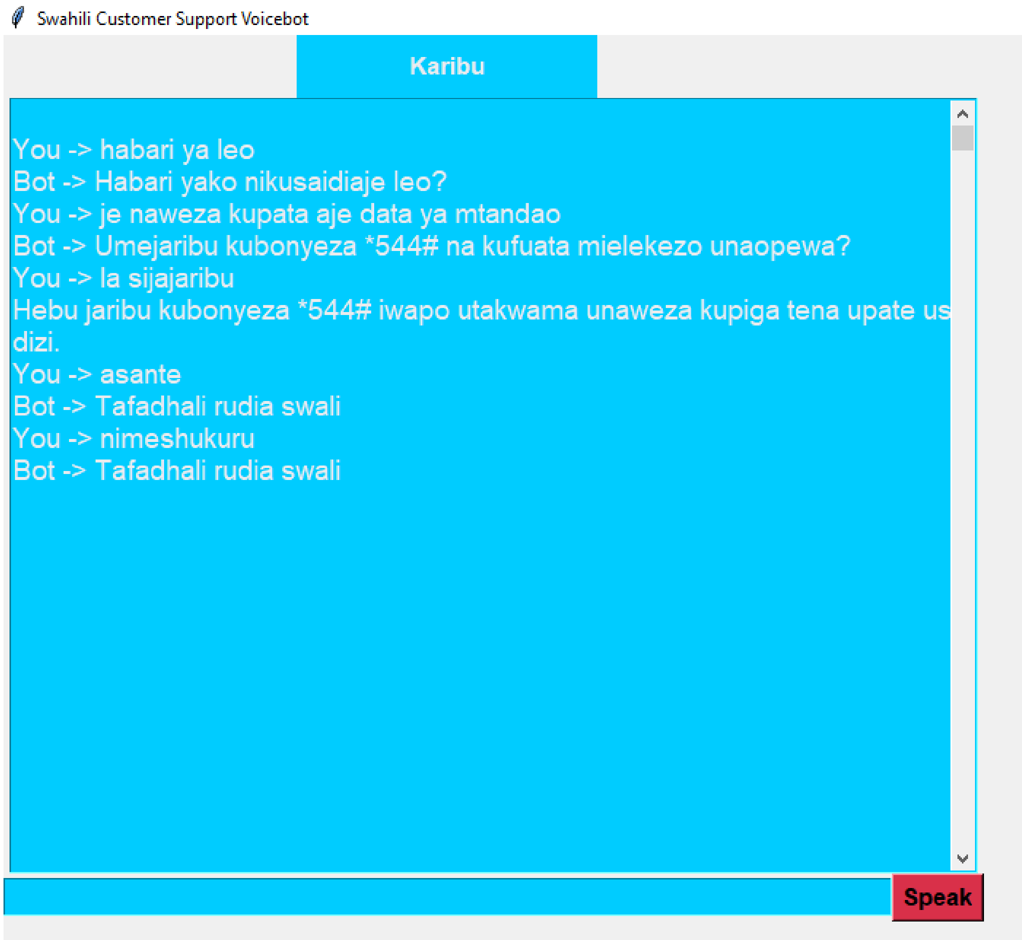

Figure 10: Sample chatbot output

4.4 Model Evaluation

Evaluation metrics for dialogue systems are often similar to those for text generation tasks such as question answering, machine translation, and summarization. For instance, as discussed earlier in section 2.6, the BERTScore leverages the pre-trained context embedding that includes cosine similarity calculations between outcomes and reference responses. Additionally, BERTScore has been evidenced to correlate with human judgement at system and sentence levels providing better evaluation for text generations especially in a domain that highly utilizes synonyms. However, unlike BLEURT, BERTScore is a lot slower that called for the evaluation of the model through Google Collaboratory hence utilizing the application's GPU.

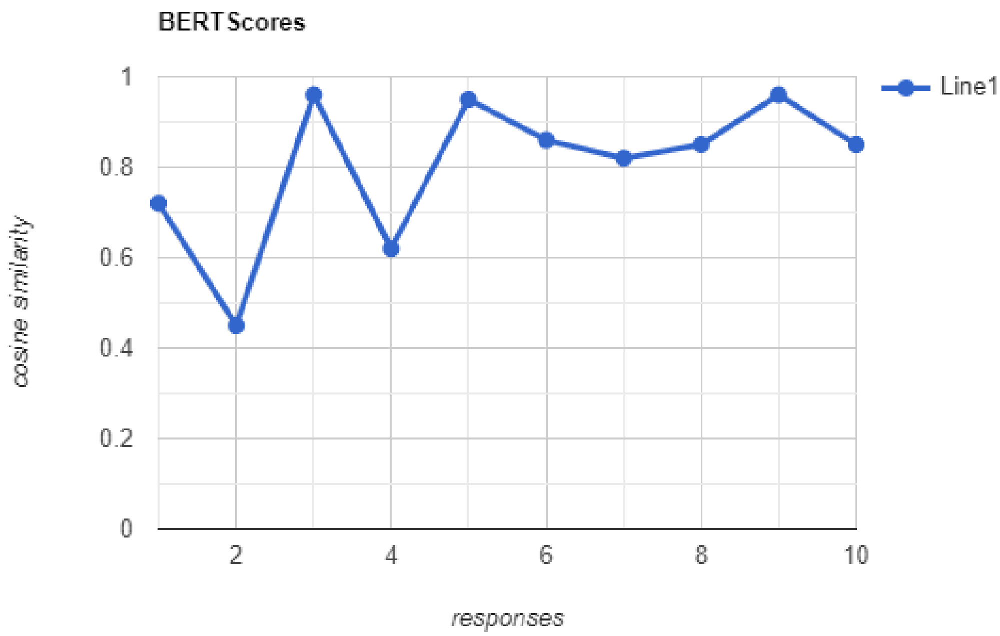

Figure 11: BERTScore results

5.0 SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

5.1 Summary of Findings

Objective 1: To collect and analyse NLP data for the Swahili language

This study set out to collect and analyse NLP data for the Swahili language to reduce the current gap in the availability of Swahili data for speech and text NLP tasks. Such ensures that future researchers have free and accessible data that they can use to solve various real-life problems, for instance, that related to conversational AI and question answering using the Swahili language. NLP data for Swahili language collected and analysed during this project were categorized into two: speech data and text data. While the speech data was used for the ASR model, the text data were adopted within the chatbot models for conversational flow. This research obtained its text data from Safaricom's FAQs section on the organization's website. As indicated in section 3.3 this research opted for the question and response conversational flow hence the selection of FAQ datasets. Additionally, a majority of the open-source FAQ datasets were open domain and included heavy technological terms limiting their effective adoption in this research. While the collected text data were originally in English they were translated to Swahili using Google translate resulting in a Swahili FAQ dataset with 88 manually created intents and 1000 pattern and response pairs that take a question and answer format. The collected speech data, on the other hand, included in-house corpus with one male and two female speakers uttering several Swahili words equivalent to 3hours of domain-specific speech data.

Objective 2: To develop a model for Swahili voicebots

This research set out to contribute to the survival and development of cultural and linguistic heritage of Swahili by developing an application that uses Swahili reducing the need for users to learn a new language and abandoning native languages while trying to access various applications. This study obtained a final Swahili voicebot model with three distinct sections: the ASR, the chatbot, and the speech synthesizer.

Objective 3: To test the validity of the proposed model

This study anticipated that the resultant voicebot model will inform and guide the design of similar such voicebots for other domains such as healthcare or insurance. Owing to such, this research set out to test the validity of the developed Swahili voicebot model for use in customer support or replication for other domains.

5.2 Conclusion

Research Question 1: What are the best approaches to the modelling of a Swahili voicebot?

Typically, there exist 3 approaches to build voicebots: 3rd party off the shelf solutions, no-code solutions, for example Google's DialogFlow or IBM Watson, and customized solutions. This research identified that the best approach to the modelling of a Swahili voicebot includes the use of customized solutions despite having used a no-code solution for the speech synthesis section.

Research Question 2: What are the best methods to test the validity the proposed model?

BERTscore is an evaluation metric utilized in the testing of the efficiency of text generation systems such as chatbots. This study established BERTscore as the most appropriate method in the testing of the voicebot's validity as it best correlates with human judgement due to its focus on computing semantic similarity between tokens of reference and system responses.

5.3 Limitations

The primary challenge faced during this research was the lack of domain specific data that would have allowed the developed models fit into the customer support domain. This limitation was however, resolved through creation of datasets from the Safaricom FAQs section. Additionally, this research was also limited by computational power and time of the machine used during creation of the ASR. Storing and access of audio files proved complex resulting in frequent crashing of the personal computer during training. This limitation was resolved through the merging of this research's acoustic model with that obtained from the Kenyan Languages Corpus (KenCorpus). This step resulted in a better acoustic model capturing more voice variants ensuring the system better understands a wider array of Swahili accents and intonations. Lastly, the challenge of computational time was also experienced during the development of a customized speech synthesizer. This limitation was resolved with the adoption of a no-code solution obtained from the Google TTS API.

5.4 Future Works

The work presented in this paper supplement previous foundations spearheading future research in conversational AI. This paper has illustrated and reviewed some of the technologies applicable in Conversational AI on low-resourced languages highlighting shortcomings of these technologies and equally providing insight applicable in the path towards adoption of conversational AI for other low-resourced languages and their testing in real-world scenarios. While previous research reviewed in this paper have evidenced various disparate results, future work can be accomplished to ensure all the identified components are merged into a single hybrid conversational AI prototype capable of performing exceptionally well for low-resourced languages. While Conversational AI in fields such as healthcare and education have in the recent past evidenced significant research, applications in fields such as customer support can be better developed by combining key tenets witnessed in other fields. Finally, NLP often requires very extensive data for effective training, manipulation, and adoption to varied case scenarios. Therefore, future work will need to prioritize data collection and application for the Swahili language prompting their availability for manipulation in the development of NLP applications for Swahili. Additionally, there should be focus on building speech synthesis models that can effectively simulate voices from low-resourced language speakers.

5.5 Recommendations

Chatbots have since become an integral part of applications today with significant efforts targeting their replication of most human behaviours. In most cases, chatbots are introduced to solve non-stringent customer problems with unsatisfied customers being transferred to human agents. This research offers a concise picture of the approaches deployable in the development of voicebots for low-resourced languages which can be further manipulated and improved. However, from this research it is quite evident that conversational AI agents are yet to pass the Turing test. This research has put several technologies towards development of voicebots for low-resourced languages into perspective providing a platform for further innovation and exploration particularly for speech synthesis for low-resourced languages which highly contributed to the developed Swahili voicebot model failing to pass the Turing test.

References

Adamopoulou, E., & Moussiades, L. (2020). Chatbots: History, technology, and applications. *Machine Learning with Applications*, *2*, 100006.

Ahmadinejad, B. (2019). The impact of customer satisfaction on word of mouth marketing (Case study: Bamilo online store). *SCIREA Journal of Management*, *3*(2), 40-52.

Assenmacher, D., Clever, L., Frischlich, L., Quandt, T., Trautmann, H., & Grimme, C. (2020). Demystifying social bots: On the intelligence of automated social media actors. *Social Media+ Society*, *6*(3), 2056305120939264.

Ayanouz, S., Abdelhakim, B. A., & Benhmed, M. (2020, March). A smart chatbot architecture based NLP and machine learning for health care assistance. In *Proceedings of the 3rd International Conference on Networking, Information Systems & Security* (pp. 1-6).

Benkerzaz, S., Elmir, Y., & Dennai, A. (2019). A study on automatic speech recognition. *Journal of Information Technology Review*, *10*(3), 77-85.

Bisong, E. (2020). Cloud Run: Google Cloud Text to Speech API. Retrieved from https://towardsdatascience.com/cloud-run-google-cloud-text-to-speech-api-dff308665c70

Campbell, S., Greenwood, M., Prior, S., Shearer, T., Walkem, K., Young, S., ... & Walker, K. (2020). Purposive sampling: complex or simple? Research case examples. *Journal of Research in Nursing*, *25*(8), 652-661.

Clinciu, M., Eshghi, A., & Hastie, H. (2021). A Study of Automatic Metrics for the Evaluation of Natural Language Explanations. *arXiv preprint arXiv:2103.08545*.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gelas, H., Besacier, L., & Pellegrino, F. (2012). Developments of Swahili resources for an automatic speech recognition system. In *Spoken Language Technologies for Under-Resourced Languages*.

Ibzan, E., Balarabe, F., & Jakada, B. (2016). Consumer satisfaction and repurchase intentions. *Developing country studies*, *6*(2), 96-100.

Jain, M., Kumar, P., Kota, R., & Patel, S. N. (2018, June). Evaluating and informing the design of chatbots. In *Proceedings of the 2018 Designing Interactive Systems Conference* (pp. 895-906).

Kaggle (2021). https://www.kaggle.com/

Karakanta, A., Dehdari, J., & van Genabith, J. (2018). Neural machine translation for low-resource languages without parallel corpora. *Machine Translation*, *32*(1), 167-189.

Lamere, P., Kwok, P., Gouvea, E., Raj, B., Singh, R., Walker, W., ... & Wolf, P. (2003, April). The CMU SPHINX-4 Speech Recognition System. In *Ieee intl. conf. on acoustics, speech and signal processing (icassp 2003), hong kong* (Vol. 1, pp. 2-5).

Lowe, R., Noseworthy, M., Serban, I. V., Angelard-Gontier, N., Bengio, Y., & Pineau, J. (2017). Towards an automatic turing test: Learning to evaluate dialogue responses. *arXiv preprint arXiv:1708.07149*.

Magueresse, A., Carles, V., & Heetderks, E. (2020). Low-resource Languages: A Review of Past Work and Future Challenges. *arXiv preprint arXiv:2006.07264*.

Nadarzynski, T., Miles, O., Cowie, A., & Ridge, D. (2019). Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study. *Digital Health, 5, 205520761987180.*

Nyongesa, G., Omieno, K., and Otanga, D. (2020). Artificial intelligence chatbot adoption framework for real-time customer support in Kenya. *International Journal of Scientific Research in Computer Science Engineering and Information Technology, 6*(6), 100-117.

Oruganti, S.C. (2020). Virtual Bank Assistance: An Ai Based Voice Bot For Better Banking. *International Journal of Research, 9*(1), 177-183

Paikari, E., & Van Der Hoek, A. (2018, May). A framework for understanding chatbots and their future. In *2018 IEEE/ACM 11th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE)* (pp. 13-16). IEEE.

Palinkas, L. A., Horwitz, S. M., Green, C. A., Wisdom, J. P., Duan, N., & Hoagwood, K. (2015). Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Administration and policy in mental health and mental health services research*, *42*(5), 533-544.

Panesar, K. (2020). Conversational artificial intelligence-demystifying statistical vs linguistic NLP solutions. *Journal of Computer-Assisted Linguistic Research*, *4*(1), 47-79.

Radhakrishnan, J., & Chattopadhyay, M. (2020, December). Determinants and Barriers of Artificial Intelligence Adoption–A Literature Review. In *International Working Conference on Transfer and Diffusion of IT* (pp. 89-99). Springer, Cham.

Reddick, C. G., Enriquez, R., Harris, R. J., & Sharma, B. (2020). Determinants of broadband access and affordability: An analysis of a community survey on the digital divide. *Cities*, *106*, 102904.

Rodríguez Cardona, D., Werth, O., Schönborn, S., & Breitner, M. H. (2019). A mixed methods analysis of the adoption and diffusion of Chatbot Technology in the German insurance sector. *Twenty-fifth Americas Conference on Information Systems, Cancun.*

Schoonenboom, J., & Johnson, R. B. (2017). How to construct a mixed methods research design. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *69*(2), 107-131.

Sellam, T., Das, D., & Parikh, A. P. (2020). BLEURT: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Smutny, P., & Schreiberova, P. (2020). *Chatbots for learning: A review of educational chatbots for the Facebook Messenger. Computers & Education, 103862.*

Shikali, C. S., & Mokhosi, R. (2020). Enhancing African low-resource languages: Swahili data for language modelling. *Data in brief*, *31*, 105951.

Surendran, A., Murali, R., & Babu, R. K. (2020). *Conversational AI-A Retrieval Based Chatbot* (No. 4020). EasyChair.

Svenningsson, N., & Faraon, M. (2019, December). Artificial intelligence in conversational agents: A study of factors related to perceived humanness in chatbots. In *Proceedings of the 2019 2nd Artificial Intelligence and Cloud Computing Conference* (pp. 151-161).

Tripathy, J. P. (2013). Secondary data analysis: Ethical issues and challenges. *Iranian journal of public health*, *42*(12), 1478.

Waqas, M., Hamzah, Z. L. B., & Salleh, N. A. M. (2020). Customer experience: a systematic literature review and consumer culture theory-based conceptualisation. *Management Review Quarterly*, 1-42.

Wu, T., Zhao, L., & Zhang, Q. (2020, December). Research on Speech Synthesis Technology Based on Rhythm Embedding. In *Journal of Physics: Conference Series* (Vol. 1693, No. 1, p. 012127). IOP Publishing.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:190*