**UNIVERSITY OF DAR ES SALAAM**

**RESEARCH PROPOSAL FOR PARTIAL FULFILLMENT OF MASTER OF SCIENCE**

**DEGREE IN DATA SCIENCE BY COURSEWORK AND DISSERTATION.**

| **Name of Candidate:** | Vicent Wilson W |
| --- | --- |
| **Name of Supervisor:** | Dr. Hellen Maziku |
| **Department /College:** | College of Information and Communication Technology |
| **Proposed Degree:** | Master of Science in Data Science |
| **Proposed Title:** | Enhancing Topic Classification of Kiswahili Text Questions using Deep Learning Methods with Word2Vec and GloVe Techniques. |

# CHAPTER ONE INTRODUCTION

## General Introduction

Kiswahili is a language spoken by a significant number of people in the Eastern and Central African region, and it is also considered as a low resource Language[(Shikali et al., 2019)](https://www.zotero.org/google-docs/?GIYxWM).The language has a common Bantu noun classes, where noun classes have two prefixes joined to a nominal stem, one for singular and other for plural.In verbs, affixes are used to indicate different grammatical relations including subject, object, tense, aspect, and mood.The affixes linked to nouns, verbs, and their dependent words with meaning-bearing can be used to infer syntactic and functional information[(Ng’Ang’A, n.d.)](https://www.zotero.org/google-docs/?iNS0Mv).

Question Answering (QA) system refers to an automated system that is capable of understanding natural language questions and providing relevant and accurate answers.The QA system substantially depends on Question Classification(QC) for allocating labels to a question into a category that represents a specific answer or topic type.The performance of the QA system depends on the accuracy of QC.The QC is not bounded to the English language only, but also other languages including Chinese[(Miao et al., 2018)](https://www.zotero.org/google-docs/?bkW6a2), Arabic[(Hasan & Zakaria, 2005)](https://www.zotero.org/google-docs/?xEn85G), and Albania[(Trandafili et al., 2018)](https://www.zotero.org/google-docs/?UdccWA).QA systems consist of two types of questions,factoid and non-factoid questions, as described by Jayalakshmi & Sheshasaayee (2015). Factoid questions require concise and straightforward answers that pertain to specific entities like places, individuals, or organizations. On the other hand, non-factoid questions demand detailed and explanatory responses, often requiring human intervention, such as response teams.

SMS platforms provide a means to develop Question Answering systems,enabling customers to ask questions and receive accurate responses from specialized response teams organized according to different topics.UNICEF-Tanzania employs an SMS platform called U-Report to implement a Kiswahili Question Answering system.This system allows them to gather data from adolescents and young individuals regarding their diverse perspectives on various subjects.Using the platform, they engage with adolescents by posing questions and responding to inquiries across different topics.In the specified system, the process of categorizing incoming questions is carried out manually, which is resource-intensive, resulting in delayed responses (UNICEF-Tanzania, 2019).However, by leveraging deep learning techniques for topic-based text classification and employing word embedding to represent Kiswahili textual data, we can develop an efficient solution.This approach allows us to train a deep learning model specifically designed for the Kiswahili question-answering (QA) system.By adopting this approach, we can enhance efficiency, significantly improve response speed, and minimize the need for direct human involvement. Moreover, employing deep learning methods and word embedding enables us to swiftly and accurately analyze the topics of Kiswahili text questions.

Pre-processing and text representation techniques play a critical role in the text classification process. It is important to employ an efficient text representation technique that is user-friendly and yields superior outcomes. Among the widely utilized methods for text representation are word embedding techniques like Word2Vec and GloVe [(C. Wang et al., 2020)](https://www.zotero.org/google-docs/?uCLiSY). These techniques encode words into compact vector representations within a continuous space, enabling effective analysis and processing of textual data.

## Statement of the problem

Currently, the manual process is used to classify text questions in Kiswahili QA systems on SMS platforms. This manual approach leads to inaccurately categorized questions, resulting in delayed responses from the relevant teams. To address this issue and enhance the effectiveness of customer support, I propose the development of an automated Kiswahili text question classification model. Accurate and timely responses are essential for customer satisfaction, yet there is a notable absence of automated Kiswahili text classification techniques in existing literature. This study aims to bridge this gap by utilizing Deep learning methods, Word2Vec , and Glove techniques to classify Kiswahili text questions, ultimately determining the most effective approach.

## Objectives

## Main Objective

The main objective is to enhance the classification of Kiswahili text questions into specific topics using deep learning methods, and word embedding techniques, specifically Word2Vec and Glove.

### Specific Objectives

1. To improve the Kiswahili common datasets for pre-processing Kiswahili textual data, specifically tailored for the task of topic classification in Kiswahili text question.
2. To train deep learning classifiers using word embedding techniques, specifically Word2Vec and Glove.
3. To evaluate the performance of the trained Deep learning classifier algorithms using metrics such as accuracy, precision, recall and F1-score.

## Research Questions

To meet the specific research objectives, the study will aim to answer the following questions:

1. What are pre-processing resources and performance impact of pre-processing steps when preparing Kiswahili textual data for Deep learning classification task?
2. What method of word embedding form of text representation between Word2Vec and GloVe may be utilized to find the best classifier for Kiswahili textual datasets?
3. What is the performance of Deep learning classifiers on the Kiswahili textual dataset when using Word2Vec and GloVe techniques?

## Significance of the study

The previous study highlighted the lack of research conducted on topic classification for Kiswahili text questions using deep learning methods and word embedding techniques in the existing body of knowledge. It emphasized the unavailability of essential resources such as pre-processed datasets specific to Kiswahili, labeled datasets for topic classification, and a dedicated corpus for machine learning tasks in Kiswahili.

The proposed study on topic classification of Kiswahili text questions using deep learning methods with Word2Vec and GloVe techniques aims to address these gaps and further contribute to the field of Natural Language Processing (NLP) for Kiswahili.

Firstly, this study will contribute by providing enhanced and accessible pre-processing datasets that include common stop words, typos, slangs, and a generated corpus specifically tailored for Kiswahili. These datasets will serve as valuable resources in the pre-processing phase and text representation of Kiswahili textual data for various NLP tasks.

Secondly, the research will propose appropriate pre-processing steps that are relevant for deep learning classification tasks involving Kiswahili textual data. This will help establish best practices and guidelines for effectively processing and preparing Kiswahili text data for deep learning models.

Lastly, the study will offer a comparative evaluation of the performance of deep learning algorithms when applied to the classification of Kiswahili text questions using different word embedding techniques, specifically Word2Vec and GloVe. This comparative analysis will provide insights into the strengths and weaknesses of these techniques for topic classification in Kiswahili and help researchers and practitioners make informed decisions when choosing word embedding methods for similar NLP tasks. By conducting this study, it aims to contribute to the advancement of NLP research and applications specifically tailored to Kiswahili, ultimately enhancing the understanding and capabilities in processing, analyzing, and classifying Kiswahili text questions.

## Scope of the study

This research will cover the train and test Deep learning Classifier algorithms which are Convolutional neural networks(CNN), Recurrent neural networks(RNN), transformer methods. It will further use the Kiswahili text questions as training and testing dataset and Word2Vec and GloVe techniques as word embedding techniques, which are forms of text representation techniques. Topics to be labeled in a dataset will be Health, Menstrual Hygiene, WASH, Nutrition, Education, HIV/AIDS, Violence Against Children (VAC), Corona, and U-Report. These topics will be used as training and testing datasets.

# 

# CHAPTER TWO LITERATURE REVIEW

## Introduction

Any research builds on the existing knowledge in a particular field. This chapter reviews various prior literatures on question answering systems and natural language processing. The review is done to analyze different strategies that will be utilized in generating a deep learning model that resolves text classification in Kiswahili. A summary of question classification, topic analysis, Kiswahili language structure and text representation of data in computation processing will be briefly reviewed in this chapter. Then the chapter will discuss various deep learning methods for classifying text data. Finally, relevant and related empirical works will be reviewed and the knowledge gap will be identified.

## Question Answering (QA) System

Question Answering (QA) is a task in Natural Language Processing (NLP) that involves the interpretation of questions and the provision of accurate responses based on those questions. The response is typically a short piece of text selected from a larger body of text. Due to advancements in NLP and Machine Learning, which have led to significant enhancements in QA systems, and the increased accessibility to the general public through mobile applications like SMS, QA has experienced substantial growth [(Madabushi & Lee, n.d.)](https://www.zotero.org/google-docs/?zOfPBH).

In information retrieval, QA systems utilize a pre-organized database or a set of texts written in natural language to automatically respond to questions posed by humans in a natural language format.

In other words, QA systems enable users to ask questions and receive answers using natural language queries. These systems can be seen as more advanced versions of data retrieval methods [(Calijorne Soares & Parreiras, 2020)](https://www.zotero.org/google-docs/?3NoXq2).

Ensuring the quality of the questions asked is crucial, as it impacts the effectiveness of the question answering system. Quality control plays a significant role in various applications of information retrieval and Natural Language Processing (NLP) in the future. Machine learning techniques effectively handle the automatic classification of text documents by leveraging labeled training documents for learning purposes [(Jayalakshmi & Sheshasaayee, 2015)](https://www.zotero.org/google-docs/?4fMBiT).

## Question Classification

The classification of questions based on their expected answer or provided topic is a crucial aspect of QA systems. In order to successfully provide accurate answers, it is essential to understand the intention behind the question. The question classification (QC) greatly impacts the overall effectiveness of QA systems. While some QA systems may not employ QC, it has been discovered that integrating QC can significantly enhance the performance of QA systems [(Madabushi & Lee, n.d.)](https://www.zotero.org/google-docs/?Jbygf7).

The primary objective of QC in QA systems is to assign appropriate labels to questions based on the predicted type of answer. Current research on question classification often focuses on employing statistical approaches using machine learning classifiers such as Support Vector Machine (SVM) and Artificial Neural Network (ANN) [(Sangodiah et al., 2005)](https://www.zotero.org/google-docs/?LMPamH).

As QA systems are transitioning from a traditional document retrieval perspective to an information retrieval standpoint, QCs are utilized as integral components within QA systems. This eliminates the need for users to sift through lengthy ranked lists of documents to find answers to their questions, saving them time and effort. QCs have significance not only in QA but also in information retrieval (IR) [(Sangodiah et al., 2005)](https://www.zotero.org/google-docs/?SZ234f).

### Classifier Algorithms for Question Classification

Text classification is a method used to determine the category of an unlabeled text by utilizing predefined topic categories. Numerous research studies have been conducted across various languages, providing different suggestions for the most effective classifiers in natural language processing (NLP) tasks.The table below shows the performance of classical machine learning algorithms for text classification for kiswahili language as it was conducted by [(Masua & Masasi, 2020)](https://www.zotero.org/google-docs/?RWDwKK):-

| Algorithm | Without Preprocessing | | All Pre-Processing Steps | | | |
| --- | --- | --- | --- | --- | --- | --- |
| Accuracy | F1 score | Accuracy | F1 Score | Accuracy Improvement (%) | F1 Score Improvement (%) |
| TF-IDF | | | | | | |
| DT | 0.6611 | 0.6628 | 0.7317 | 0.7326 | **10.68** | 10.53 |
| KNN | 0.7394 | 0.7819 | 0.7450 | 0.8002 | 0.75 | 2.35 |
| MNB | 0.5820 | 0.5286 | 0.6364 | 0.5922 | 9.34 | **12.03** |
| RF | 0.7875 | 0.7819 | 0.8012 | 0.7966 | 1.74 | 1.88 |
| SGD | 0.8052 | 0.8012 | **0.8226** | **0.8183** | 2.16 | 2.13 |
| SVM | 0.8104 | 0.8091 | 0.8123 | 0.8100 | 0.23 | 0.11 |
| Doc2Vec | | | | | | |
| SGD | 0.7654 | 0.7733 | 0.8188 | 0.8218 | 6.98 | 6.27 |
| SVM | 0.7901 | 0.7941 | 0.8192 | 0.8284 | 3.68 | 4.32 |
| RF | 0.6846 | 0.7193 | 0.7363 | 0.7889 | **7.55** | **9.68** |
| LSTM-Uni | 0.8150 | 0.8247 | **0.8379** | **0.8476** | 2.77 | 2.78 |
| LSTM-Bi | 0.8070 | 0.8024 | 0.8363 | 0.8375 | 3.63 | 4.37 |

Table 2.1 Text Classification for kiswahili Language

The table below also shows the performance of different study for different machine learning algorithms for different languages:

| Author Name & Title | Classifier | Evaluation | datasets | language |
| --- | --- | --- | --- | --- |
| (Hasan & Zakaria, 2016) – Question Classification Using Support Vector Machine and Pattern Matching | support vector machine (SVM) with n-gram and WordNet | F-measure at 87.25 with the 2-gram feature yields | 200 questions about Hadith from Sahih Al Bukhari | Arabic |
| (D. Zhang & Lee, 2003) - Question Classification using Support Vector Machines | K-Nearest Neighbors (KNN) | Accuracy average of 70.52% | 5,500 labeled questions were randomly divided into five training datasets of 1,000, 2,000, 3,000, 4,000, and 5,500 questions each. The TREC10 QA track has 500 labeled questions in the testing dataset. | English |
| Naïve Bayes | Accuracy average 66.02% |
| Decision Tree | Accuracy average 76.85% |
| Sparse Network of  Winnows (SNoW). | Accuracy average 69.96% |
| SVM | Accuracy average 79.26% |
| (Sangodiah et al., 2015) - Question Classification Using Statistical Approach: A Complete Review | SVM with Radial Biased Function (RBF) - kernel function | Accuracy of 84.12% | 4394 questions in the two-layered question taxonomy, with 6 rough categories and 65 fine categories. | Chinese |
| (Rajvanshi & Chowdhary, 2017) - Comparison of SVM and Naïve Bayes Text  Classification Algorithms using WEKA | Naïve Bayes | Accuracy of 85.53% | The Car Dataset has 1728 total instances with 6 characteristics (buying capacity, maintenance, number of doors, seating Capacity, boot space, safety and class) | English |
| SVM with RBF - kernel function | Accuracy of 94.21% |
| (Sheshasaayee & Thailambal, 2017) - Comparison of Classification Algorithms in Text  Mining | Random Forest | Accuracy of 88% | 500 samples in a sample dataset with a Positive or Negative class. | English |
| Naïve Bayes | Accuracy of 90.2% |
| SVM | Accuracy of 97.4% |
| (Miao et al., 2018) - Chinese News Text Classification Based on Machine learning algorithm | KNN | F-measure at 0.919 takes 19.60s | News dataset of 9331 News datasets with nine categories (agriculture, economy, history, environment, politics, sports, art, computer, space) | Chinese |
| Naive Bayesian | F-measure at 0.920 takes 3.42s |
| SVM with Term Frequency × Inverse Document Frequency (TF×IDF) | F-measure at 0.957 takes 484.75s |
| (Trandafili, Kote, et al., 2018) - Performance Evaluation of Text Categorization  Algorithms Using an Albanian Corpus | Simple Logistic | Accuracy of 80.7% | Document dataset of 800 questions datasets with twenty categories (Animals, Art, Astronomy, Biology, Charity, Chemistry, Culture, Curiosities, Economy, Environment, Fashion, Food, History, Literature, Medicine, Politics, Religion, Sport, Technology and Tourism) | Albanian |
| Naive Bayesian | Accuracy of 82.7% |
| KNN | Accuracy of 44.7% |
| Decision Tree | Accuracy of 66.6% |
| SVM | Accuracy of 82.9% |
| Random Forest | Accuracy of 75.9% |
| Simple Logistic | Accuracy of 80.7% |

Table 2.2 Text Classification Works on Different Languages

Different QC work on classifying textual data using ML classifier algorithms that have been done in other languages shows that ML classifiers for textual data are Random Forest, SVM, SGD, Decision Tree, KNN, ANN and Naïve Bayes. Classifiers that are commonly used are SVM and Naïve Bayes with a combination of pre-processing steps like punctuation removal, lowercasing, stop word removal, misspelling checking and slang correction. QC studies also show that TF×IDF weighting methods or Doc2vec text representation technique can improve positively performance of Classifiers in both accuracy and F1 score.

1. **Convolutional Neural Networks (CNN)**

A convolutional Neural Networks classifier is a type of Deep learning classifier that uses convolutional layers to extract local features from input text, followed by pooling layers to capture the most salient features. The extracted features are then fed into fully connected layers for classification [(Kim, 2014)](https://www.zotero.org/google-docs/?zkN2ES).

1. **Recurrent Neural Networks (RNN):**

Recurrent Neural Network is a type of deep learning architecture that is designed to capture sequential dependencies in data, making them particularly suitable for processing sequential and text-based data.RNNs can be used to model the sequential nature of text by considering the order of words or characters in a sentence. They process the input text sequentially, word by word or character by character, while updating hidden states at each step. This allows RNNs to retain information about the previous words or characters in the sequence and utilize it for classification [(Zaman & Mishu, 2017)](https://www.zotero.org/google-docs/?XKkhRG).

1. **Transformer**

A Transformer is a revolutionary deep learning architecture that is based on the concept of self-attention, allowing it to capture contextual dependencies between words or tokens in a sequence without relying on recurrent or convolutional operations. It avoids sequential processing and enables parallelization, making it highly efficient for both training and inference.The key components of the Transformer architecture include, Encoder, Multi-Head Attention, Positional Encoding, Feed-Forward Neural Networks, Classification Layer.Transformers have achieved state-of-the-art results in various text classification tasks. Pretrained transformer models like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pretrained Transformer) have been widely used and fine-tuned on specific classification tasks, demonstrating their effectiveness in capturing contextual information and achieving high classification accuracy [(Ashish, 2017)](https://www.zotero.org/google-docs/?dTlb2c).

### Artificial Neural Networks

The utilization of artificial neural networks (ANN) in natural language processing (NLP) involves various techniques such as establishing subject-oriented databases for text data, conducting frequency analysis, developing subject-oriented dictionaries, and performing tokenization and digital vectorization of texts. The preprocessing steps for text data, including input importation, word partitioning, dictionary construction, index translation, and the creation of training and test samples, are consistent across the ANN layers. However, it is important to note that while NLP methods like tokenization and vector representation primarily focus on English, other languages like Chinese, Russian, Arabic, and more encounter additional challenges in their implementation [(Rogachev et al., 2021)](https://www.zotero.org/google-docs/?W2pm1g).

Machine and deep learning algorithms, utilizing artificial neural networks (ANN), have gained popularity for evaluating text content through sentiment and semantic analysis. These approaches have been widely adopted by researchers in various domains, including social media analysis and health information websites. Deep neural networks have the capability to capture complex structures and non-linear transformations, allowing them to model high-level abstractions and effectively reduce the dimensionality of data across multiple processing layers [(Jelodar et al., 2020)](https://www.zotero.org/google-docs/?cngqG7).The mathematical expression representing each LSTM cell can be defined as:

ƒ*t* = (*W*ƒɀɀ*t* – 1 + *WƒxXt* + bƒ) ……………………………... 1

*it* = (*Wiɀɀt* – 1 + *WixXt* + c*i*) ………………………………... 2

*ot* = (*Woɀɀt* – 1 + *W0xXt* + b*0*) ……………………………… 3

Equations 1 to 3 determine the forget (ƒt), input (it), and output (ot) gates for each LSTM cell, respectively.In this context, Xt represents the input vector, and ɀt corresponds to the hidden output. The weight matrix and bias term are denoted by Wand b, respectively, while σ refers to the sigmoid function. The forget gate plays a role in deciding which previous information in the cell state should be disregarded within an LSTM layer. On the other hand, the input gate governs the retention of new information within the memory cell. Lastly, the output gate regulates the amount of information that is revealed from the internal memory cell [(Jelodar et al., 2020)](https://www.zotero.org/google-docs/?yoiSjE).

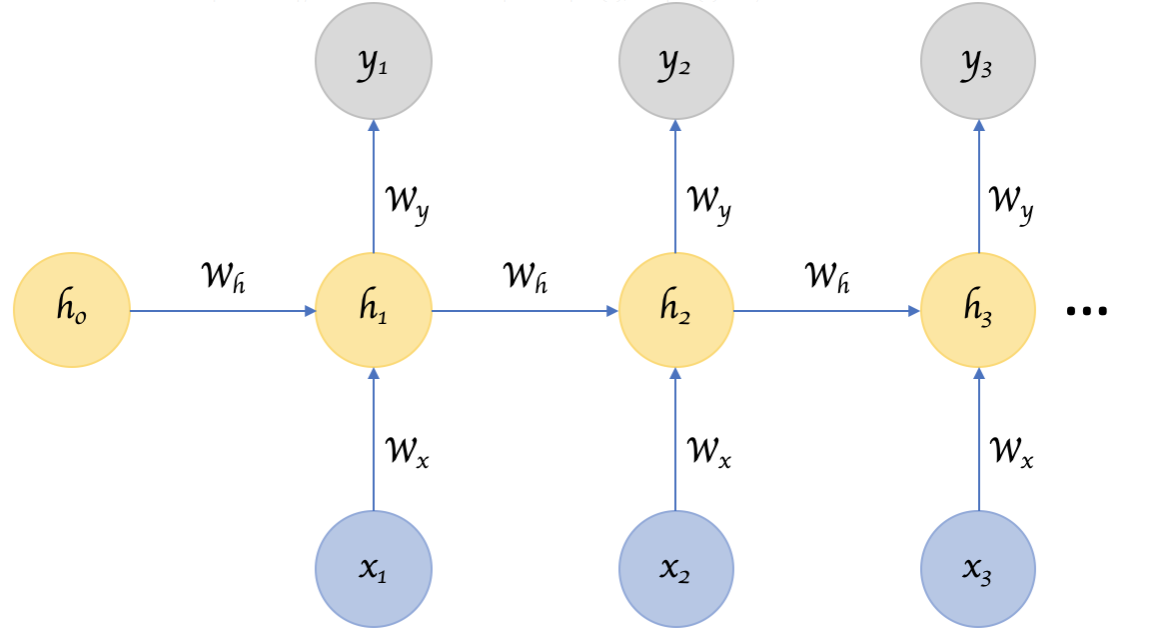


Figure 2.1 Unidirectional LSTM [(Cui et al., n.d.)](https://www.zotero.org/google-docs/?tEcffp)

Two variants of RNN LSTM, namely Bidirectional and Unidirectional LSTM, exist. When employing the Bidirectional LSTM, inputs are processed in two directions: one from the past to the future and another from the future to the past. This approach differs from the Unidirectional LSTM since it incorporates an LSTM layer that operates in reverse to retain information from the future. By combining the two hidden states, Bidirectional LSTM enables the preservation of information from both the past and the future at any given time. Figure 2.1 illustrates the processing approach of the Unidirectional LSTM. In this method, each weight, represented by "w," pertains to distinct cells from various layers. The input data layer is denoted as "x," the inner/processing layer as "h," and the output layer as "y" [(Sherstinsky, 2020)](https://www.zotero.org/google-docs/?4RoVLY).

In certain situations, relying solely on past information is insufficient for predicting the future. There are instances where it becomes necessary to have a glimpse into the future in order to rectify the past. Particularly in tasks like speech recognition and handwriting recognition, where a single component of the input can present a considerable amount of ambiguity, having knowledge of what follows next can aid in understanding the context and accurately recognizing the present. The bidirectional LSTM addresses this text processing scenario by examining the subsequent output to validate the current output. Figure 2.2 depicts the processing approach of the bidirectional LSTM, with the input data layer represented as "x," the inner processing layer from past to future denoted as "h," the inner processing layer from future to past labeled as "g," and the output layer as "y" [(Cui et al., n.d.)](https://www.zotero.org/google-docs/?Dtm2tq).

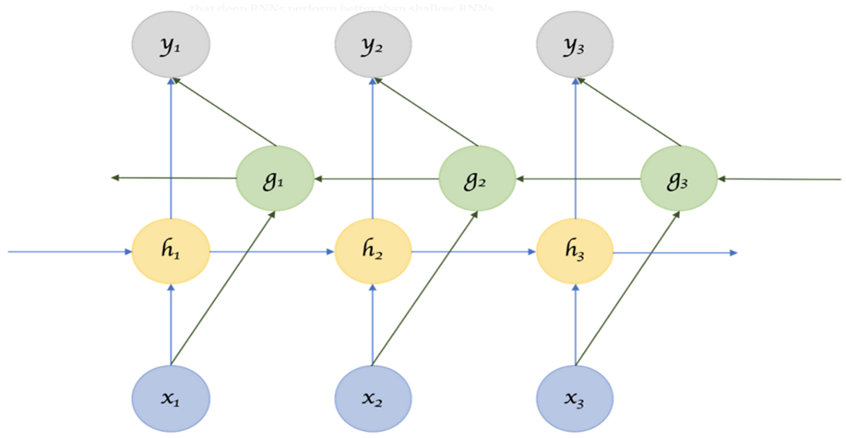


Figure 2.2 Bidirectional LSTM (Cui et al., 2018)

[(Jelodar et al., 2020)](https://www.zotero.org/google-docs/?031sJs) suggests a deep learning model for utilizing Long Short-Term Memory (LSTM) for the classification of sentiment in COVID-19-related comments. This model demonstrated superior performance compared to several other widely recognized machine learning techniques.

## Topic Analysis

Topic analysis is a method in machine learning that facilitates the organization and comprehension of extensive text data by assigning tags or categories to each text based on its subject or theme. By employing topic analysis, one can reveal patterns and semantic structures within individual texts. Topic modeling and topic classification are two machine learning techniques used for topic analysis [(*Chauhan2017.Pdf*, n.d.)](https://www.zotero.org/google-docs/?Q2LRXL).

Topic modeling is an unsupervised machine learning technique that can identify patterns and group similar utterances without relying on predefined topic labels or training data. This method is straightforward to implement but suffers from one drawback: it tends to be less accurate [(Steyvers & Griffiths, 2014)](https://www.zotero.org/google-docs/?ph5xeP). On the other hand, topic classification is a supervised machine learning technique that necessitates prior knowledge of the themes present in the text before conducting the analysis. This approach requires labeled data to train a topic classifier. Despite the additional step involved, topic classifiers yield long-term benefits and offer higher precision compared to clustering approaches [(Yin et al., 2012)](https://www.zotero.org/google-docs/?8pw8mw).

## Kiswahili Language Structure

Kiswahili utilizes a noun class system in the style of Bantu languages, where two prefixes are added to the nominal stem to indicate singular and plural forms. Verbs in Kiswahili employ affixes to convey various grammatical relations such as subject, object, tense, aspect, and mood. The concordia agreement system requires that nouns and other sentence elements agree in class and number with the verb in a sentence. Adjectives, possessive pronouns, and demonstratives also exhibit agreement with the noun they modify in terms of class and number. The affixes attached to nouns, verbs, and their dependent words carry meaningful information, which can be used to derive syntactic and functional details. This information can then be employed for semantic clustering and classification purposes [(Ng’Ang’A, n.d.)](https://www.zotero.org/google-docs/?tdSc1X).

In Kiswahili, the subject typically precedes the verb, which is then followed by the object. Noun modifiers are positioned after the noun, similar to how adverb modifiers appear after the verb they modify. The arrangement of words in the immediate context of a target word holds valuable information regarding the target word's semantic and syntactic characteristics, especially in languages with a fixed word order. This enhances the availability of contextual information and contributes to the acquisition of semantic clustering for Kiswahili words [(Ng’Ang’A, n.d.)](https://www.zotero.org/google-docs/?VfjzMh).

In Kiswahili, the focus of information structure lies more in the context of discourse where specific pieces of propositional information are conveyed, rather than solely on the interpretation of words and sentences within particular conversational settings. This implies that every sentence in Kiswahili is inherently linked to its information structure and cannot exist without it[(Ndung’u & Box, 2015)](https://www.zotero.org/google-docs/?61XA4L).

In most simple cases, the fundamental sentence structure of any language follows either a subject-verb-object (SVO) or subject-verb (SV) formula. However, as sentences become more complex, including some relatively simple ones, languages can differ in terms of word order, the use of unique words, tense, and context. Kiswahili stands out from other languages due to its basic syllable structure, which lacks consonant clusters, final consonants, and adds a vowel to loanwords that end with a consonant.

To enhance the performance of natural language processing (NLP) tasks, high-quality vectors are employed for text representation. Traditional language models construct text representation by considering contextual word information extracted from extensive corpora based on the language's structure. In Kiswahili, word representation vectors can be generated using sub-word units such as characters, n-grams, morphemes, and syllable-like units. Additionally, the syllable units within the structure of the Kiswahili language can be utilized to add semantic meaning to word representations [(Shikali et al., 2019)](https://www.zotero.org/google-docs/?84B31P).

## Related work

A paper written by [(Masua & Masasi, 2020)](https://www.zotero.org/google-docs/?0j6LDf) on enhancing text pre-processing for Swahili language.the study provides valuable resources for swahili language which can help to improve the accuracy and efficiency of machine learning models text.in which they developed a dataset for common swahili stop-words, slangs and typos with equivalent proper words.The datasets were stored in Comma Separated Value file format with 8-bit Unicode Transformation Format, and are publicly accessible for use in machine learning pre-processing stages.

A paper written by [(Masua & Masasi, 2020)](https://www.zotero.org/google-docs/?SVbQq8) on classification of Kiswahili text questions according to topics using TF-IDF and Doc2Vec Techniques.leveraging the classical machine learning algorithms for classifying Kiswahili text questions into specific topics. The classifiers used were, Random Forest, SVM, SGD, Decision Tree, KNN, RNNs with LSTM and Naïve Bayes. the performance of this study After performing all pre-processing steps showed, SGD algorithm outperform other algorithms with accuracy of 82.26% and f1-score of 81.83% by using TF-IDF feature selection approach. Also, the LSTM-Unidirectional model outperforms other models with accuracy of 82.26% and f1-score of 81.83% by using Doc2Vec feature selection approach.

A paper written by [(H. Wang & Li, 2022)](https://www.zotero.org/google-docs/?5rjNfm) on Chinese news text classification based on convolutional neural network (CNN). leveraging a combined-convolutional neural network text classification model based on word2vec and improved TF-IDF for Chinese news text classification. The model uses a six-layer CNN architecture and a text representation method based on word vector and weight combination. The model is evaluated on the Thucnews dataset and achieves a test accuracy of 97.56%, which is better than traditional Text-RNN model, traditional Text-CNN model, and word2vec-CNN model.

A paper written by [(Aydoğan & Karci, 2020)](https://www.zotero.org/google-docs/?OJ4f4z) on improving the accuracy using pre-trained word embeddings on deep neural networks for Turkish text classification. by leveraging on enhancing the accuracy of Turkish text classification through the utilization of pre-trained word embeddings in deep neural networks.The authors explore the effectiveness of different pre-trained word embedding models, such as Word2Vec and GloVe, in the context of Turkish text classification.The findings demonstrate that integrating pre-trained word embeddings into deep neural networks results in improved accuracy for Turkish text classification. The use of pre-trained word embeddings captures the semantic relationships between words more effectively, leading to enhanced classification performance. The study highlights the importance of utilizing pre-trained word embeddings as a valuable resource for enhancing the accuracy of Turkish text classification tasks.

## Research gap

The majority of research in Question Classification has focused on using machine learning classifier algorithms to classify textual data in languages such as Swahili, English, Arabic, Albania, and Chinese. Since languages differ in their structure, the same applies to Kiswahili language. Thus, there is a necessity to classify Kiswahili textual data based on specific topics and assess the performance of trained classifiers. To accomplish the classification of Kiswahili textual data, it is crucial to explore necessary pre-processing steps like removing punctuation marks and stop words, substituting slangs and typos, and understand how these steps impact the classifier's performance. Additionally, it is essential to investigate commonly used word embedding techniques in NLP tasks for other languages, determine their impact on the accuracy of classifiers using an adequate Kiswahili corpus, and identify the classifier that outperforms the others

# CHAPTER THREE RESEARCH METHODOLOGY

## Research Design

This research is of an experimental nature, where certain sets of variables are kept constant while another set of variables is measured as the focus of the experiment. The study involves conducting experiments using various pre-processing techniques with Word2Vec and GloVe word embedding techniques, which serve as methods for text representation. The objective is to apply different Deep learning classifier algorithms on classifying constant Kiswahili question message dataset . The performance of each trained Deep learning classifier algorithm is then evaluated to determine the best classifier.

## Research Approach

### Pre-processing

Data pre-processing plays a pivotal role in refining the dataset for subsequent analysis. In this research, the focus will be on enhancing data quality and ensuring its suitability for analysis.

Initially, I will address issues related to data quality, including identifying and rectifying incomplete, unreliable, inaccurate, or irrelevant data segments. To achieve this, I will employ established pre-processing techniques, such as the Inspection method, to detect and rectify unexpected, incorrect, and inconsistent data. Any anomalies detected will be corrected or removed to ensure data integrity.

Furthermore, linguistic aspects, such as punctuation removal and lowercasing, will be applied to standardize the text. While typo correction is a necessary step, it's worth noting that the dataset already includes a set of appropriate words for correcting common Kiswahili typos made by writers.Similarly, for slang correction, the dataset already contains a collection of proper words for addressing Kiswahili slang used by writers. This means there is more access to valuable resources for refining the text.

After these initial pre-processing steps, I will perform a dataset verification process to validate the effectiveness of the applied techniques in improving data quality.

Finally, standard NLP (Natural Language Processing) pre-processing steps, such as stop word removal, vectorization, and normalization, will be applied to prepare the dataset for subsequent training and testing phases.

### Training Deep learning Classifier Algorithms

The training process for the algorithms will begin by identifying Deep learning classifier algorithms that are commonly utilized in textual data for text classification. An extensive review of previous studies on evaluating Deep learning classifiers in textual data classification will be conducted. This review will provide insights into the performance of each classifier, allowing the selection of classifiers that have demonstrated superior performance. To develop a better understanding of the Kiswahili language, a Kiswahili Corpus will be generated, specifically tailored to suit the selected Deep learning classifiers. The pre-processing dataset in Kiswahili will undergo tokenization and vectorization using Word2Vec and GloVe word embedding techniques, which are text representation methods that convert text into numerical values. Subsequently, the pre-processed dataset will be split into training and test sets. The final step will involve training the selected classifier algorithms using the training dataset and fine-tuning the parameters for each classifier.

### Evaluating performance of Trained Classifiers

The evaluation of the training classifier algorithms will involve assessing various parameters specific to each classifier. The performance of each classifier will be measured by calculating accuracy, precision and recall scores on a separate set of future data.In this study, the training classifier algorithms will be assessed using a robust cross-validation approach to comprehensively evaluate their performance. The primary metrics for evaluation will include accuracy, recall, and precision. To achieve this, the confusion matrix method will be employed, which will allow to calculate precision and recall scores for each training classifier. This method provides insights into how well the classifiers can correctly classify positive instances while minimizing false positives.the evaluation process will be repeated to ensure robustness, and it will incorporate both word embedding techniques. This ensures that the classifiers' performance is assessed across different representations of the data.ultimately, the training classifiers with the most optimal hyperparameters will be chosen based on their performance in terms of accuracy, recall, and precision. This selection process will help to identify the most effective models for the topic classification task.

## Study Area

The research will be conducted in Tanzania, which is one of the East and Central African countries with a significant population of Kiswahili speakers compared to other countries in the region. The organization UNICEF-Tanzania focuses on gathering data from adolescents and young individuals in Tanzania through an SMS platform. This platform allows them to ask questions or provide answers on various topics, resulting in a substantial volume of Kiswahili text questions. Since its establishment in April 2017, the platform has attracted over 220,000 registered adolescents and young people.

## Tools and Materials

This study will utilize a variety of materials, including but not limited to: -

| **S/No** | **Material** | **Tool description** |
| --- | --- | --- |
| 1 | portable computer | for computation |
| 2 | software | JetBrains PyCharm for the analysis |
| 3 | modem | for Internet access to retrieve various literature materials |
| 4 | interview | for capturing quantitative data |
| 5 | Programming Languages | Python version 3.10.4, JavaScript ECMAScript 2023 |
| 6 | Data formats | JavaScript Object Notation (JSON), Comma Separated Value (CSV), Excel (xls, xlsx) |
| 7 | Virtual Environment | Pipenv |
| 8 | Web Development Frameworks | Django 3.1.6, StreamLit |
| 9 | Version control | Git |
| 10 | Libraries | NumPy, Pandas, JupyterLab, Scikit-learn, Gensim, Keras matplotlib |

Table 3.1 tools and materials to be used in this study

## Data Collection

The data for this study will be gathered from the UNICEF-Tanzania U-Report SMS platform, which involves Kiswahili text questions posed by U-Reporters to the UNICEF response team. The questions cover a wide range of topics including health, menstrual hygiene, WASH, nutrition, education, HIV/AIDS, Violence Against Children (VAC), Corona, and U-Report. The data spans from 2019 to 2022. U-Report is a social messaging platform in Tanzania that primarily operates through SMS, aiming to empower and engage young people to express their concerns, contribute to citizen-led development, respond to emergencies, and effect positive change. The platform utilizes the RapidPro framework, an open-source system that facilitates the processing of SMS messages for multiple mobile networks. Through RapidPro, U-Report collects both desired and undesired Kiswahili text questions, enabling real-time data collection and communication with the target users (UNICEF-Tanzania, 2019).

This study will employ a combination of primary and secondary data collection methods. Secondary data will be acquired from various sources such as Kiswahili reports, articles, newspapers, social media, and SMS platforms. This data will be used to compile datasets for the corpus, typos, slang, and stop words. Primary data collection will involve accessing the U-Report SMS platform and gathering SMS messages sent by U-Reporters. Additionally, documents will be reviewed to gather existing slang data from published or unpublished sources.

### Document Review

As part of the secondary data collection, this study will review multiple documents from diverse sources including research papers, academic articles, journals, open-data platforms, and relevant websites. The purpose of this document review is to gain insights into different techniques used for analyzing and classifying textual data. It aims to gather information about the approaches adopted by other researchers, their methodologies, and the findings they obtained. Additionally, the review will aid in identifying existing Deep Learning classifiers and selecting suitable models that can be customized to align with the specific context of this research.

### Interview

Once the secondary data has been obtained, an interview will be conducted with the UNICEF - Tanzania Team responsible for U-Partnership to gather additional information about the collection of SMS datasets from the U-Report SMS Platform. The interview will focus on inquiries such as the number of response teams operating within the U-Report SMS system, the required categories for classification, and the specific topics that can be categorized based on the collected SMS data. Additionally, interviews will be conducted with Kiswahili language experts at TUKI (Taasisi ya Ukuzaji Kiswahili) of UDSM. These interviews will serve the purpose of gaining a deeper understanding of the structure of the Kiswahili language, as well as reviewing and validating the primary and secondary data collected, including Kiswahili slang, Kiswahili stop words, and common Kiswahili typos.

## Data Analysis

In this research, the application of Exploratory Data Analysis (EDA) will be employed to conduct initial investigations on the collected data. The purpose is to uncover patterns, derive meaningful insights, identify anomalies, gain a comprehensive understanding of the data, and provide summary statistics and graphical representations. The initial data analysis will specifically focus on the datasets containing texts sent to and received from U-Reporters during the period from 2019 to 2022. To facilitate this process, a Python virtual environment will be established. Python scripts will be developed to handle the large JSON files associated with each month, enabling them to be split into smaller chunks. These chunks will be saved in the Comma Separated Value (CSV) format, ensuring ease of use and effective data manipulation.

## Ethical Consideration

To address the ethical considerations associated with this study, research clearance letters will be obtained from UDSM and also the research clearance letters will be provided to three organizations involved in data collection: UNICEF, Y4C hub, and IKS. The potential respondents will be informed about the benefits of the study and will be requested to participate voluntarily throughout the data collection process. They will be fully informed and assured that this research is a necessary component of academic requirements and will be utilized to develop a tool that employs a Kiswahili deep learning model for the classification of Kiswahili questions asked by young individuals in Tanzania through the UNICEF U-Report SMS platform

## REFERENCES

[Ashish, V. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, I.](https://www.zotero.org/google-docs/?N9iOgn)<https://doi.org/10.1109/EICT.2017.8275236>

[Aydoğan, M., & Karci, A. (2020). Improving the accuracy using pre-trained word embeddings on deep neural networks for Turkish text classification. *Physica A: Statistical Mechanics and Its Applications*, *541*, 123288. https://doi.org/10.1016/j.physa.2019.123288](https://www.zotero.org/google-docs/?N9iOgn)

[Calijorne Soares, M. A., & Parreiras, F. S. (2020). A literature review on question answering techniques, paradigms and systems. *Journal of King Saud University - Computer and Information Sciences*, *32*(6), 635–646. https://doi.org/10.1016/j.jksuci.2018.08.005](https://www.zotero.org/google-docs/?N9iOgn)

[*Chauhan2017.pdf*. (n.d.).](https://www.zotero.org/google-docs/?N9iOgn)

[Cui, Z., Ke, R., Pu, Z., & Wang, Y. (n.d.). *Stacked Bidirectional and Unidirectional LSTM Recurrent Neural Network for Network-wide Traffic Speed Prediction*.](https://www.zotero.org/google-docs/?N9iOgn)

[Hasan, A. M., & Zakaria, L. Q. (2005). QUESTION CLASSIFICATION USING SUPPORT VECTOR MACHINE AND PATTERN MATCHING. . *. Vol.*](https://www.zotero.org/google-docs/?N9iOgn)

[Jayalakshmi, S., & Sheshasaayee, A. (2015). Question Classification: A Review of State-of-the-Art Algorithms and Approaches. *Indian Journal of Science and Technology*, *8*(29). https://doi.org/10.17485/ijst/2015/v8i29/87466](https://www.zotero.org/google-docs/?N9iOgn)

[Jelodar, H., Wang, Y., Orji, R., & Huang, S. (2020). Deep Sentiment Classification and Topic Discovery on Novel Coronavirus or COVID-19 Online Discussions: NLP Using LSTM Recurrent Neural Network Approach. *IEEE Journal of Biomedical and Health Informatics*, *24*(10), 2733–2742. https://doi.org/10.1109/JBHI.2020.3001216](https://www.zotero.org/google-docs/?N9iOgn)

[Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1746–1751. https://doi.org/10.3115/v1/D14-1181](https://www.zotero.org/google-docs/?N9iOgn)

[Madabushi, H. T., & Lee, M. (n.d.). *High Accuracy Rule-based Question Classification using Question Syntax and Semantics*.](https://www.zotero.org/google-docs/?N9iOgn)

[Miao, F., Zhang, P., Jin, L., & Wu, H. (2018). Chinese News Text Classification Based on Machine Learning Algorithm. *2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*, 48–51. https://doi.org/10.1109/IHMSC.2018.10117](https://www.zotero.org/google-docs/?N9iOgn)

[Ndung’u, N. M., & Box, P. O. (2015). *Information Structure in Kiswahili*. *3*(3).](https://www.zotero.org/google-docs/?N9iOgn)

[Ng’Ang’A, W. (n.d.). *SEMANTIC ANALYSIS OF KISWAHILI WORDS USING THE SELF ORGANIZING MAP*.](https://www.zotero.org/google-docs/?N9iOgn)

[*Rapidpro | UNICEF Office of Innovation*. (n.d.). Retrieved May 26, 2023, from https://www.unicef.org/innovation/topics/rapidpro](https://www.zotero.org/google-docs/?N9iOgn)

[Rogachev, A., Melikhova, E., & Atamanov, G. (2021). *Building Artificial Neural Networks for NLP Analysis and Classification of Target Content:* Conference on current problems of our time: the relationship of man and society (CPT 2020), Yakutsk, Russia. https://doi.org/10.2991/assehr.k.210225.058](https://www.zotero.org/google-docs/?N9iOgn)

[Sangodiah, A., Muniandy, M., & Heng, L. E. (2005). QUESTION CLASSIFICATION USING STATISTICAL APPROACH: A COMPLETE REVIEW. . *. Vol.*](https://www.zotero.org/google-docs/?N9iOgn)

[Sherstinsky, A. (2020). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. *Physica D: Nonlinear Phenomena*, *404*, 132306. https://doi.org/10.1016/j.physd.2019.132306](https://www.zotero.org/google-docs/?N9iOgn)

[Shikali, C. S., Sijie, Z., Qihe, L., & Mokhosi, R. (2019). Better Word Representation Vectors Using Syllabic Alphabet: A Case Study of Swahili. *Applied Sciences*, *9*(18), 3648. https://doi.org/10.3390/app9183648](https://www.zotero.org/google-docs/?N9iOgn)

[Steyvers, M., & Griffiths, T. (2014). Probabilistic Topic Models. In *Handbook of Latent Semantic Analysis*. Routledge. https://doi.org/10.4324/9780203936399.ch21](https://www.zotero.org/google-docs/?N9iOgn)

[Trandafili, E., Kote, N., & Biba, M. (2018). Performance evaluation of text categorization algorithms using an Albanian Corpus. *Advances in Internet, Data & Web Technologies: The 6th International Conference on Emerging Internet, Data & Web Technologies (EIDWT-2018)*, 537–547.](https://www.zotero.org/google-docs/?N9iOgn)

[Wang, C., Nulty, P., & Lillis, D. (2020). A Comparative Study on Word Embeddings in Deep Learning for Text Classification. *Proceedings of the 4th International Conference on Natural Language Processing and Information Retrieval*, 37–46. https://doi.org/10.1145/3443279.3443304](https://www.zotero.org/google-docs/?N9iOgn)

[Wang, H., & Li, X. (2022). Chinese News Text Classification Based on Convolutional Neural Network. *Journal on Big Data*, *4*(1), 41–60. https://doi.org/10.32604/jbd.2022.027717](https://www.zotero.org/google-docs/?N9iOgn)

[Yin, Z., Cao, L., Gu, Q., & Han, J. (2012). Latent Community Topic Analysis: Integration of Community Discovery with Topic Modeling. *ACM Transactions on Intelligent Systems and Technology*, *3*(4), 1–21. https://doi.org/10.1145/2337542.2337548](https://www.zotero.org/google-docs/?N9iOgn)

[Zaman, M. M. A., & Mishu, S. Z. (2017). Convolutional recurrent neural network for question answering. *2017 3rd International Conference on Electrical Information and Communication Technology (EICT)*, 1–6. https://doi.org/10.1109/EICT.2017.8275236](https://www.zotero.org/google-docs/?N9iOgn)

## OTHER RELEVANT INFORMATION

## Financial Arrangement

## Sponsorship

The funding for this research is self-sponsored.

## Proposed Budget

The Proposed budget for this research is as shown in Table 5.1

**Table 5.1: Research Costs**

| **SN** | **ITEM** | **DESCRIPTION** | **TOTAL(Tsh.)** |
| --- | --- | --- | --- |
| **1.** | External Hard disk- 1TB | Backup storage for data collected from subjects | 150,000/= |
| **2** | Stationary | Printing,Photocopying and binding documents | 500,000/= |
| **3.** | Data Collection | Traveling costs to data sources | 300,000/= |
| **4.** | Internet Bundle | Internet Bundle costs | 500,000/= |
| **5.** | Dissertation Production | As stipulated in UDSM postgraduate prospectus | 350,000/= |
| **6.** | Laptop | For Model development, data collection and analysis | 2000,000/= |
| **7.** | Mi-fi Router | For Internet connection | 200,000/= |
| **8.** | Other Cost | includes meals | 500,000/= |
| **Total** |  |  | **4,500,000/=** |

## Duration of the Study

Table 5.2 below is the Gantt chart that shows the duration of study in months from May 2023 to July 2024

**Table 5.2: Time Schedule for Research**

| **Activity** | | | | | **2023** | | | | | | | | **2024** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | | | **May** | **June** | **July** | **Aug** | **Sept** | **Oct** | **Nov** | **Dec** | **Jan** | **Feb** | **March** | **April** | **May** | **June** | **July** |
| **Literature Review** | | | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Research Proposal Writing and Submission** | | | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Data Collection** | | | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Data Preprocessing, Cleaning, and Analysis** | | | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Model development** | | | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Results and Interpretation** | | | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Report Writing** | | | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Report Submission** | | | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

Candidate’s Name**:**…Vicent Wilson W………… Candidate’s Signature**:**………W.Vicent…..

Date………11st August 2023……………..

**Supervisors’(s) & Principal’s Comments**

Comment by Supervisor

…………………………………………………………………………………………………………

…………………………………………………………………………………………………………

Date**:**……………………… Name**:**……………………………… Signature …………………..

SUPERVISOR

Principal’s comment

…………………………………………………………………………………………………………

…………………………………………………………………………………………………………

Date**:**……………………… Name**:**……………………………… Signature**:**

PRINCIPAL