THE UNIVERSITY OF DODOMA

COLLEGE OF INFORMATICS AND VIRTUAL EDUCATION

DEPARTMENT OF INFORMATION SYSTEMS AND TECHNOLOGY

FINAL YEAR PROJECT PROGRESS REPORT.

PROJECT TITLE: SWAHILI TRANSCRIPTION TOOL.

SUPERVISOR'S NAME: MR. AGUSTINO MWOGOSI.

SUPERVISOR'S SIGNATURE:____________________________

S/NO	STUDENT'S NAME	REGISTRATION NUMBER	PHONE NUMBER	PROGRAM
1.	RIGOBERT KIATA	T/UDOM/2020/00333	0745560331	BSC.IS
2.	JULIETH SHAYO	T/UDOM/2020/07084	0672890521	BSC.IS
3.	CLEOPHAS KAJETANI	T/UDOM/2020/07095	0769800459	BSC.IS
4.	JOSEPH MPONJOLI	T/UDOM/2020/00332	0769115101	BSC.IS
5.	SULEMANI BAKARI	T/UDOM/2020/00336	0627881951	BSC.IS

Table of Contents

List of Tables and Figures.

List of Abbreviations.

ABBREVIATION	DEFINITION
AI	Artificial Intelligence
API	Application Program Interface
ASR	Automatic Speech Recognition
HMM	Hidden Markov Model
IEEE	Institute of Electrical and Electronics Engineering
LSTM	Long Short-Term Memory
MFCC	Mel-Frequency Cepral Coefficients
ML	Machine Learning
NLP	Natural Language Processing
RNN	Recurrent Neural Network
TALN	Technology Access For Life
TZ	Tanzania

CHAPTER ONE.

1. Introduction.

1.1. Project overview.

The project we are doing is based upon Machine Learning and we are aiming to develop a **swahili transcription tool** using machine learning. We will develop a model using machine learning natural language processing algorithms and finely tune the model to give the highest percentage of accuracy in transcribing the swahili language words into swahili text.

We will later deploy the model that we will develop into a web-based application using a python web application building framework, thus creating an acoustic model and a batch streaming application of the model.

Developing this tool, we are expecting to have the first swahili transcription tool created in Tanzania, as swahili language is originally from the bantu people of the eastern Africa decent. With unlike the rest of the transcription tools that have been developed by different developers from universities and colleges around Europe and Australia which have no great knowledge of the swahili language compared to the indigenous people of East Africa.

1.2. Problem Statement.

Artificial Intelligence is preferred as a technology that solves human problems with ease, it uses computer sciences with programming knowledge for computers to understand simple instructions from human intelligence and later mimic the human behavior at a faster and more accurate level (Robert Fay, 2021). The subset of Artificial Intelligence which is Machine Learning (ML) is the leading of this advancement. This advancement has led to the development of tools that assist humans with different tasks such as transcribing and translation, with ease, accuracy and fast execution.

Natural language processing technology in the Artificial Intelligence and Machine Learning computer sciences has grown and developed very greatly in the past couple of years, as seen with its implementation in multiple technological companies through the use of chat bots and Artificial

Intelligence assistance technologies. For example; Apple having Siri and Samsung having Bixby. All of these have been made through the use of natural language processing technology in machine learning. (Ma, 2022).

Focusing on transcribing, Machine Learning has enabled people in need of this service to finally be able to acquire the resources and services at a much cheaper price and with faster

Execution of the service with little time consumption (Ranchal, 2013). Transcription tools are available in multiple language, even swahili, but with the swahili language there is a low accuracy of the AI model. We have realized that AI models, the natural language processing field struggles with accents even with the English language and so the AI models lack a bit of accuracy due to accents (Hyryn, 2020), then the swahili language also has multiple accents that vary with the indigenous people within Tanzania. This is may be cause of the fact that the data for the manipulation of the model did not originate from the swahili people parse. The existing models fail even to capture simple swahili words like "ujenzi", "maisha", "uzoefu" and lots of more other words that can frequently be heard from the swahili people especially in the swahili speaking regions in Tanzania like Dar Es Salaam (Masua, 2020).

With the Swahili transcription tool, we will have the first transcription tool built by Tanzanians and with the implementation of speech recognition optimization a normal Swahili accent will be used as a benchmark for the rest of accents to fall through, thus increasing the accuracy of this tool far greater than the ones built by Swahili scholars abroad who are not well acquainted with the Swahili language than the indigenous people of Tanzania where Swahili is largely used compared to other parts of the world.

1.3. Project Objectives.

In this section we will be evaluating the objectives that we have set for the accomplishment of our project, the goals that we are meaning to attain at the end of the entire project.

1.3.1. General Objectives.

- To develop a swahili transcription tool using machine learning.

1.3.2. Specific Objectives

i. To find, download and prepare a swahili words dataset for transcription.

ii. To develop a swahili transcription model using natural language processing algorithms.

iii. To implement the model by deploying it in a web-based application.

1.4. Justification of the Project.

1.4.1. Purpose of the Project.

This project is aimed at innovation in a technology which is our country is not familiar with and that is transcription using machine learning, of which machine learning is a subset of artificial intelligence. Transcription using machine learning is reliable and effective employing it in various fields of study such research, art & music, data science and collection activities and even in record keeping and report writing (Chen, 2019).

The data collected for this project's undertaking and foregoing are adequate and have all been collected and pre-processed for the purpose of transcription and recording and analyzation at a research and innovation level.

1.4.2. Purpose Statement.

This purpose of this project is to create a fully functional transcription tool deployed on a web-based application using a python web development framework. Using machine learning and the experience of the swahili language which is our language in Tanzania for improving and increasing the machine learning model's accuracy of transcribing.

1.4.3. Importance of the Project.

We are making this project even though it is already developed and tested by other country developers (Gelas, 2011) and the difference we are making in training and testing the speech recognition API and NLP algorithms is that we will train the model according to our own experience with the swahili language which is far greater than other countries that use the language, since Tanzania is the mother land for swahili.

CHAPTER TWO.

2. Literature Review.

There are some transcription tools that we have come across in the search for previous works done by developers based upon Machine Learning and AI, these tools are used for transcription and translation. Platforms such as happy scribe from Barcelona https://www.happyscribe.com/transcribe-swahili of which offers transcription services for the swahili language, it still hasn't reached a desired accuracy for its transcription tool, The model is estimated to have an 85% accuracy in transcribing but it is still inaccurate with the Swahili language as the people who developed it are not originally swahili speakers.

There are other tools for transcription such as vocalmatic https://vocalmatic.com/languages/transcribe-tanzanian-swahili-to-text from Toronto, Ontario in the united states of America. This tool offers transcription services but it has a low accuracy when it comes to the swahili language, the problem it encounters is still the same with the happy scribe tool which is originally being unable to recognize the swahili speech in its original accent from the roots of the language. Natural Language Processing (NLP) always struggles with understanding human language when it is spoken with a different unfamiliar accent, not only in swahili but also other languages' as well.

When it comes to swahili transcription tools and their inaccuracy mainly due to the fact that swahili text and voice datasets are very inadequately found. The swahili language is under resourced and the amount of data available for transcription that has been used previously in creating transcription tools is inaccurate compared to other languages (Hadrien Gelas, 2012).

The accuracy of any machine learning can be determined by the amount of data the model is fed to interpret and understand human instructions giving the most accurate output and predictions of the model. The data itself can not only be used for transcription alone but also translation and as computers are to be programmed to mimic human behavior and so the same challenges, obstacles that humans undergo while performing the transcription tasks should be taken into account for, The data collection, study, analysis and documentation should further more consider the fact that languages' vary between speakers and also the clarity of audio files can have an effect upon the understanding of the language for both human level understanding and machine level of understanding (Himmelmann, 2018).

CHAPTER THREE.

3. Methodology.

We have divided project methodology into three main parts according to their Objectives.

1. To create a Swahili words dataset for transcription.

   Data Collection: Collect a large amount of audio and transcription data. This data can be collected through various sources such as public datasets, user-generated content, or by conducting interviews or surveys under Speech-to-Text API, Automatic Speech Recognition (ASR) software and Web scraping tools.

   Data Annotation: Annotate the collected audio data with corresponding transcriptions. This process is known as transcription; it can be done manually or using some automatic speech recognition tools like Manual Annotation Tools, Manual Annotation Tools, Manual Annotation Tools.

   Data Cleaning: Clean and preprocess the collected data to remove any noise or inconsistencies. This may include removing background noise, correcting errors in the transcriptions, and removing duplicates by using Audio Editing Software and data cleaning scripts.

2. To create a Swahili transcription model using natural language processing algorithm.

   Data Splitting: Split the collected and cleaned data into training, validation, and test sets. The training set is used to train the machine learning model, the validation set is used to tune the model's hyper parameters, and the test set is used to evaluate the model's performance.

   Feature Extraction: Extract features from the audio data that will be used as input to the NLP model. This may include extracting spectral features such as Mel-frequency cepstral coefficients (MFCCs) or using speech-to-text API's to transcribe the audio data.

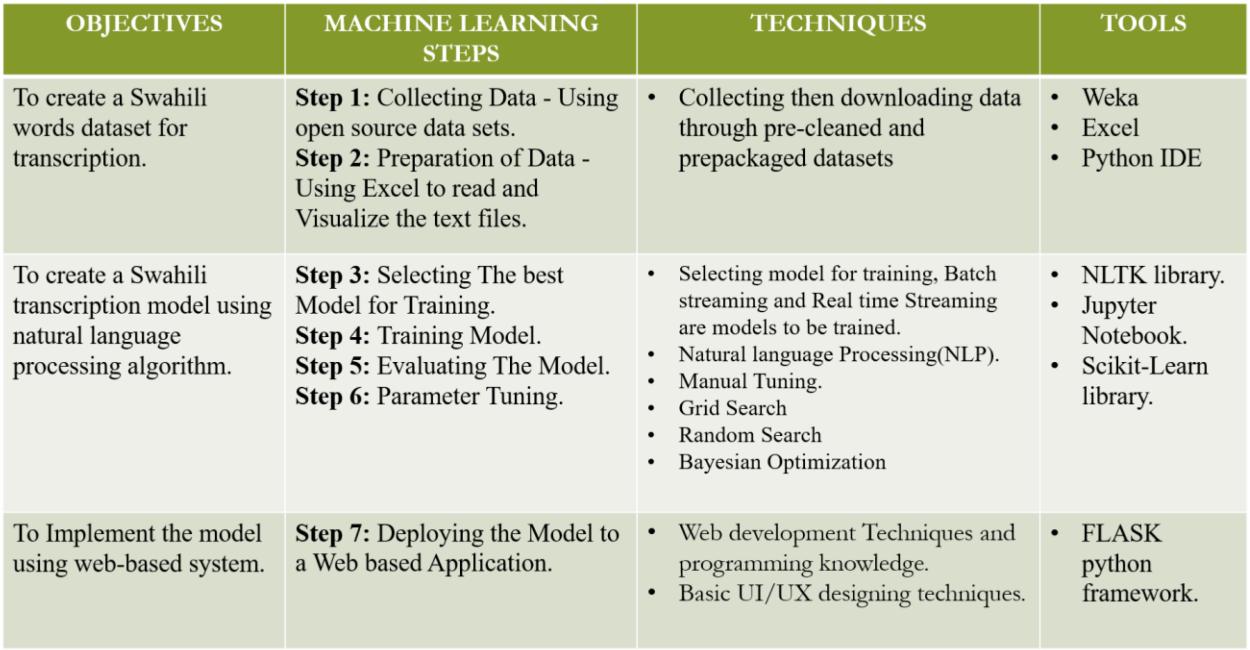
Model Selection: Select an NLP algorithm that will be used to transcribe the audio data. Some popular NLP algorithms for transcription include hidden Markov models (HMMs), recurrent neural networks (RNNs), and long short-term memory (LSTM) networks.

Model Training: Train the selected NLP algorithm on the extracted features and corresponding transcriptions. This can be done using tools such as TensorFlow, PyTorch, or Keras.

Model Evaluation: Evaluate the performance of the trained model on the test set. This can be done by comparing the model's transcriptions to the ground-truth transcriptions.

3. To Implement the model using web-based system.
   Model Deployment: Deploy the trained model in a production environment. This can be done by integrating the model into a transcription tool or by making the model available through an API under the use of frameworks like Python Flask framework

Below is a showing our project’s methodology tools and techniques aligned with the project objectives.



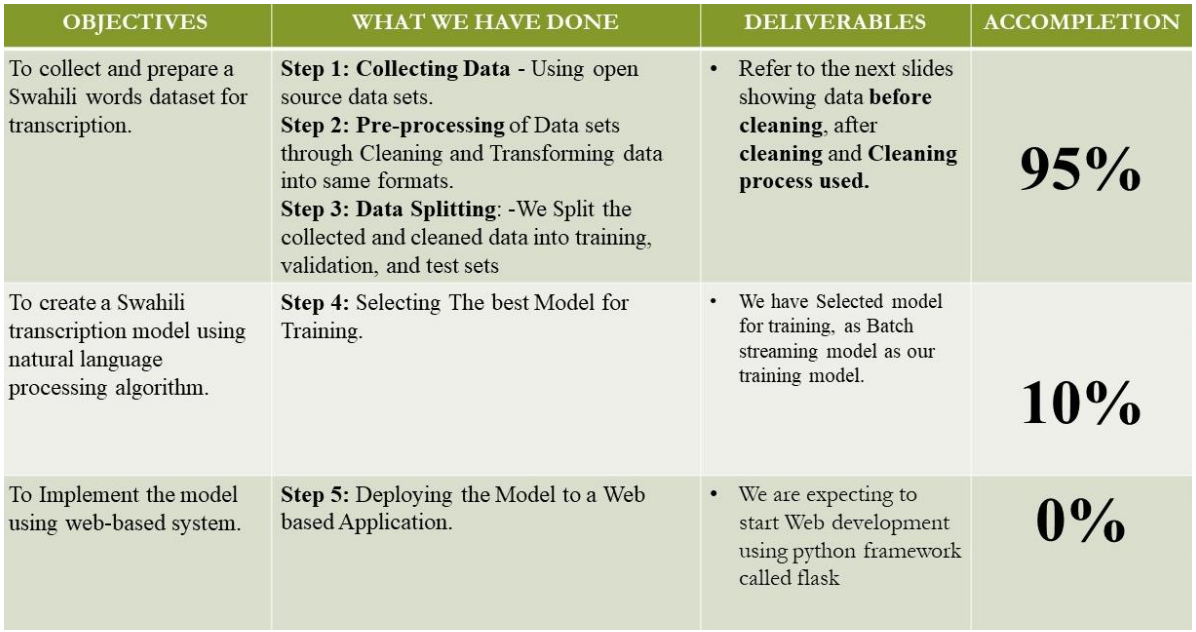
OBJECTIVES	MACHINE LEARNING STEPS	TECHNIQUES	TOOLS
To create a Swahili words dataset for transcription.	**Step 1:** Collecting Data - Using open source data sets. **Step 2:** Preparation of Data - Using Excel to read and Visualize the text files.	• Collecting then downloading data through pre-cleaned and prepackaged datasets	• Weka • Excel • Python IDE
To create a Swahili transcription model using natural language processing algorithm.	**Step 3:** Selecting The best Model for Training. **Step 4:** Training Model. **Step 5:** Evaluating The Model. **Step 6:** Parameter Tuning.	• Selecting model for training, Batch streaming and Real time Streaming are models to be trained. • Natural language Processing(NLP). • Manual Tuning. • Grid Search • Random Search • Bayesian Optimization	• NLTK library. • Jupyter Notebook. • Scikit-Learn library.
To Implement the model using web-based system.	**Step 7:** Deploying the Model to a Web based Application.	• Web development Techniques and programming knowledge. • Basic UI/UX designing techniques.	• FLASK python framework.

Table 1 Project Methodologies, tools and techniques

CHAPTER FOUR.

4. Work Done So Far.

In this chapter is the work done so far of the project, outlined after will be a summary table of all objectives and work done so far under each methodology for this project so far, and also after the work done so far is the future plans for the project after the work done in the below phase of the project.



OBJECTIVES	WHAT WE HAVE DONE	DELIVERABLES	ACCOMPLETION
To collect and prepare a Swahili words dataset for transcription.	**Step 1: Collecting Data** - Using open source data sets. **Step 2: Pre-processing** of Data sets through Cleaning and Transforming data into same formats. **Step 3: Data Splitting**: -We Split the collected and cleaned data into training, validation, and test sets	• Refer to the next slides showing data **before cleaning**, after **cleaning** and **Cleaning process used.**	**95%**
To create a Swahili transcription model using natural language processing algorithm.	**Step 4:** Selecting The best Model for Training.	• We have Selected model for training, as Batch streaming model as our training model.	**10%**
To Implement the model using web-based system.	**Step 5:** Deploying the Model to a Web based Application.	• We are expecting to start Web development using python framework called flask	**0%**

Table 2. Work done so far summary table.

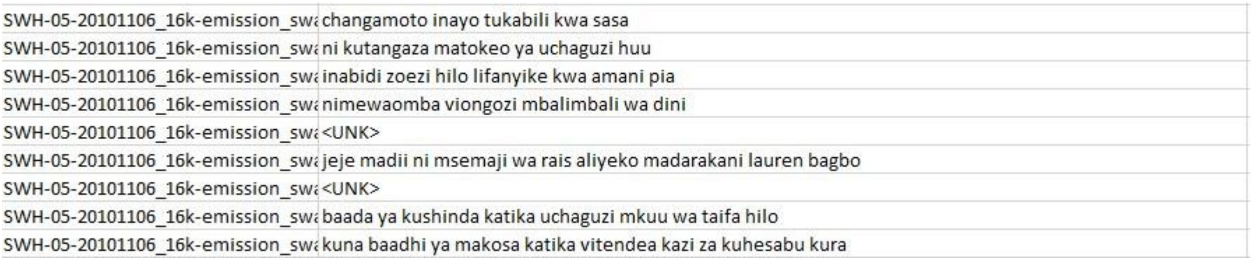
4.1. Data Collection and preparation.

1st Step in our work.

We collected data and prepared our data sets for training into our machine learning model for transcription purposes, we have collected our data set from online free repositories such as GitHub repository called https://github.com/zelalemgetahun9374/Swahili-Speech-To-Text#overview and also from https://github.com/getalp/ALFFA_PUBLIC for the audio files for the testing splitting phases of our model.

2nd Step in our work.

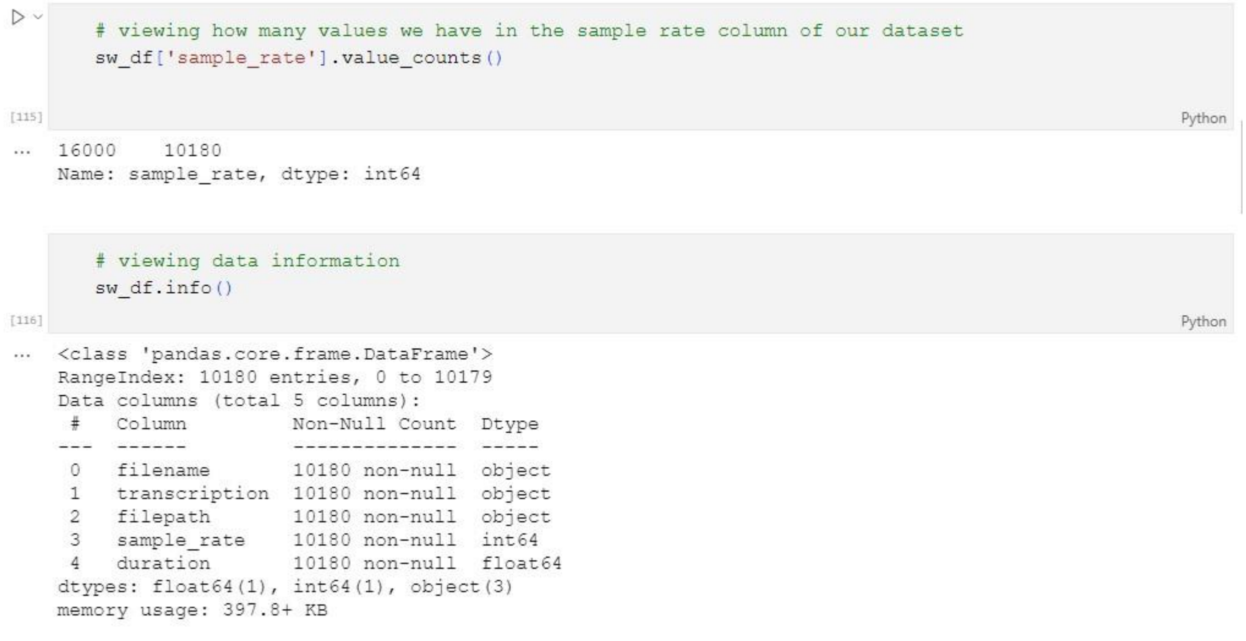
We viewed our data sets and analyzed what we required before pre-processing the data set and cleaning it for accurate model training and testing. Below is an image of our dataset before cleaning, just the raw data of 10180 instances and 5 features in the data set.



SWH-05-20101106_16k-emission_swa	changamoto inayo tukabili kwa sasa
SWH-05-20101106_16k-emission_swa	ni kutangaza matokeo ya uchaguzi huu
SWH-05-20101106_16k-emission_swa	inabidi zoezi hilo lifanyike kwa amani pia
SWH-05-20101106_16k-emission_swa	nimewaomba viongozi mbalimbali wa dini
SWH-05-20101106_16k-emission_swa	<UNK>
SWH-05-20101106_16k-emission_swa	jeje madii ni msemaji wa rais aliyeko madarakani lauren bagbo
SWH-05-20101106_16k-emission_swa	<UNK>
SWH-05-20101106_16k-emission_swa	baada ya kushinda katika uchaguzi mkuu wa taifa hilo
SWH-05-20101106_16k-emission_swa	kuna baadhi ya makosa katika vitendea kazi za kuhesabu kura

Figure 1, Data set before cleaning and pre-processing

We observed from our data set above that there are multiple instances that the collection has unknown values recorded that we need to eliminate, they were not many the count was 18 out of 10180 instances in our data set. Below is a description of the data set that we collected.



```
# viewing how many values we have in the sample rate column of our dataset
sw_df['sample_rate'].value_counts()
```

[115] Python

```
16000    10180
Name: sample_rate, dtype: int64
```

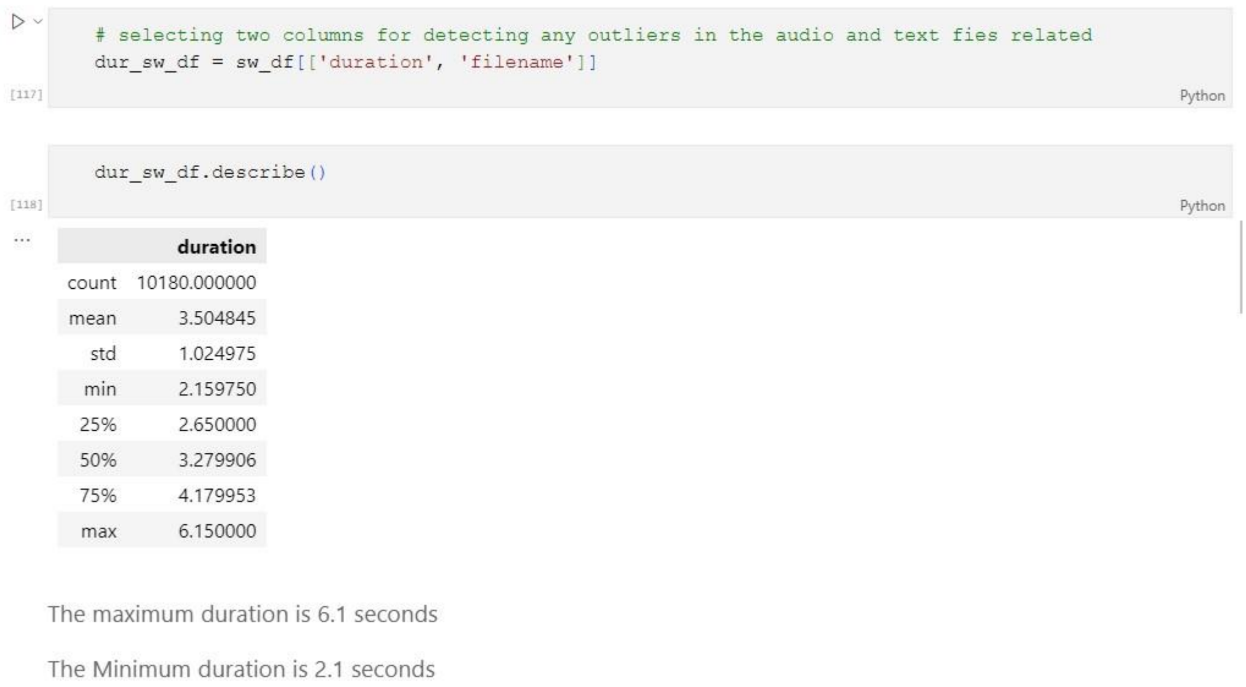
```
# viewing data information
sw_df.info()
```

[116] Python

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10180 entries, 0 to 10179
Data columns (total 5 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   filename       10180 non-null  object
 1   transcription  10180 non-null  object
 2   filepath       10180 non-null  object
 3   sample_rate    10180 non-null  int64
 4   duration       10180 non-null  float64
dtypes: float64(1), int64(1), object(3)
memory usage: 397.8+ KB
```

Figure 2; The description of our data set.

After Viewing the data we observed that our data contains 5 features and 10180 instances of data, we started cleaning the data set and checking for outliers and the figures below show the cleaning process and outliers view in a box plot. The selected feature is the duration in which the length of the transcription and audio file will determine the accuracy of the model, if the audio file is too long then the model's accuracy decreases, hence the duration of each audio file in relation to its text file has been considered and firmly adjusted to 2 seconds to 6 seconds per each audio file.



```
[117]
# selecting two columns for detecting any outliers in the audio and text fies related
dur_sw_df = sw_df[['duration', 'filename']]
```

```
[118]
dur_sw_df.describe()
```

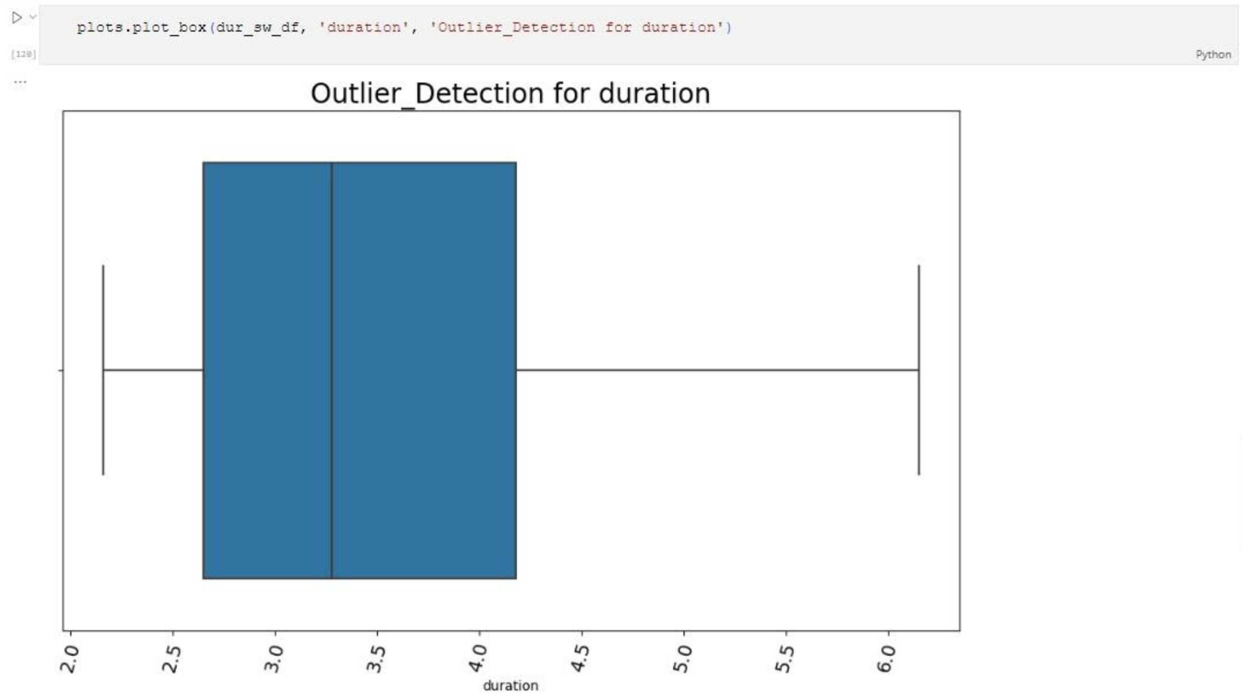
	duration
count	10180.000000
mean	3.504845
std	1.024975
min	2.159750
25%	2.650000
50%	3.279906
75%	4.179953
max	6.150000

The maximum duration is 6.1 seconds

The Minimum duration is 2.1 seconds

Figure 3; A view of the duration feature in the data set with filename feature.

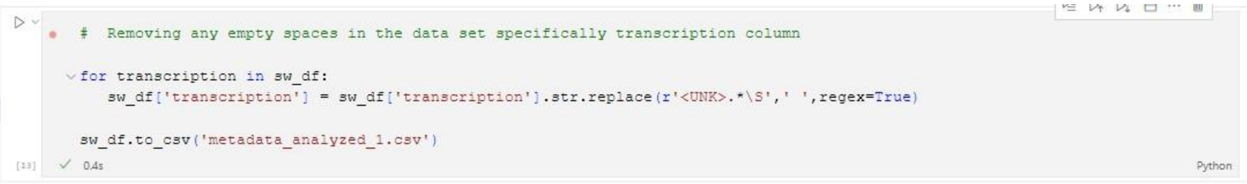
After analyzing using statistical methods in python, we observed the maximum duration in seconds, the minimum duration in seconds and the most occurring timeline in the transcribed file. Below is a view of any outliers in the duration feature of the data set.



```
plots.plot_box(dur_sw_df, 'duration', 'Outlier_Detection for duration')
```

Figure 4; Box Plot checking for outliers in the Duration feature of the Data Set.

After checking for outliers in the duration feature then we cleaned the data for every unknown entity recorded in the transcription feature of the data set because we do not need that value. Below is a figure showing how we cleaned the data sets and how we presented the data set after all the pre-processing and cleaning.



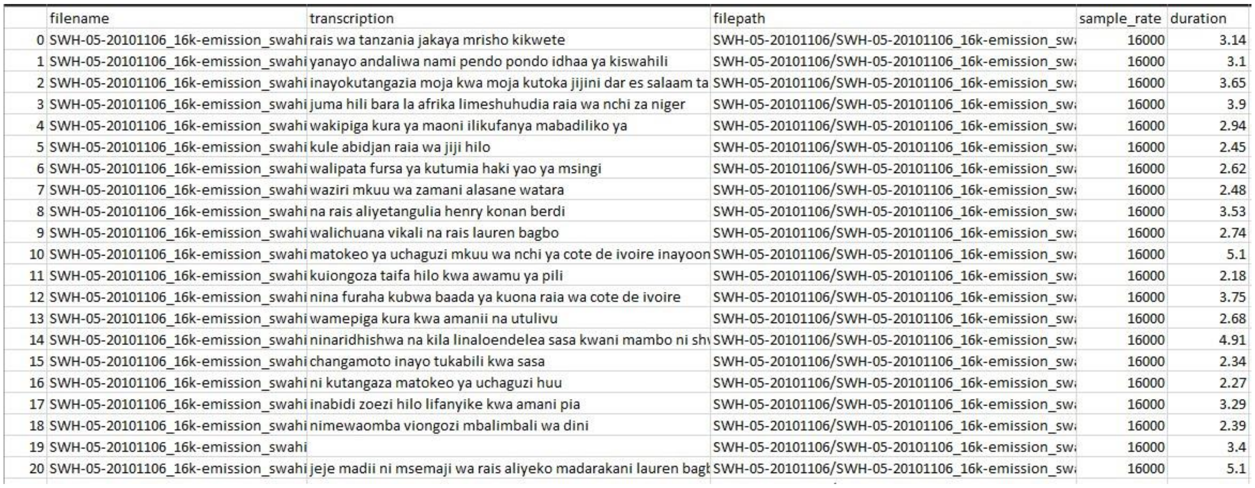
```
#  Removing any empty spaces in the data set specifically transcription column

for transcription in sw_df:
    sw_df['transcription'] = sw_df['transcription'].str.replace(r'<UNK>.*\S',' ',regex=True)

sw_df.to_csv('metadata_analyzed_1.csv')
```

Figure 5; Removing spaces in the transcription features entities.

Below is a view of the data set after cleaning and pre-processing the data set.



	filename	transcription	filepath	sample_rate	duration
0	SWH-05-20101106_16k-emission_swahi	rais wa tanzania jakaya mrisho kikwete	SWH-05-20101106/SWH-05-20101106_16k-emission_swa	16000	3.14
1	SWH-05-20101106_16k-emission_swahi	yanayo andaliwa nami pendo pondo idhaa ya kiswahili	SWH-05-20101106/SWH-05-20101106_16k-emission_swa	16000	3.1
2	SWH-05-20101106_16k-emission_swahi	inayokutangazia moja kwa moja kutoka jijini dar es salaam ta	SWH-05-20101106/SWH-05-20101106_16k-emission_swa	16000	3.65
3	SWH-05-20101106_16k-emission_swahi	juma hili bara la afrika limeshuhudia raia wa nchi za niger	SWH-05-20101106/SWH-05-20101106_16k-emission_swa	16000	3.9
4	SWH-05-20101106_16k-emission_swahi	wakipiga kura ya maoni ilikufanya mabadiliko ya	SWH-05-20101106/SWH-05-20101106_16k-emission_swa	16000	2.94
5	SWH-05-20101106_16k-emission_swahi	kule abidjan raia wa jiji hilo	SWH-05-20101106/SWH-05-20101106_16k-emission_swa	16000	2.45
6	SWH-05-20101106_16k-emission_swahi	walipata fursa ya kutumia haki yao ya msingi	SWH-05-20101106/SWH-05-20101106_16k-emission_swa	16000	2.62
7	SWH-05-20101106_16k-emission_swahi	waziri mkuu wa zamani alasane watara	SWH-05-20101106/SWH-05-20101106_16k-emission_swa	16000	2.48
8	SWH-05-20101106_16k-emission_swahi	na rais aliyetangulia henry konan berdi	SWH-05-20101106/SWH-05-20101106_16k-emission_swa	16000	3.53
9	SWH-05-20101106_16k-emission_swahi	walichuana vikali na rais lauren bagbo	SWH-05-20101106/SWH-05-20101106_16k-emission_swa	16000	2.74
10	SWH-05-20101106_16k-emission_swahi	matokeo ya uchaguzi mkuu wa nchi ya cote de ivoire inayoon	SWH-05-20101106/SWH-05-20101106_16k-emission_swa	16000	5.1
11	SWH-05-20101106_16k-emission_swahi	kuiongoza taifa hilo kwa awamu ya pili	SWH-05-20101106/SWH-05-20101106_16k-emission_swa	16000	2.18
12	SWH-05-20101106_16k-emission_swahi	nina furaha kubwa baada ya kuona raia wa cote de ivoire	SWH-05-20101106/SWH-05-20101106_16k-emission_swa	16000	3.75
13	SWH-05-20101106_16k-emission_swahi	wamepiga kura kwa amanii na utulivu	SWH-05-20101106/SWH-05-20101106_16k-emission_swa	16000	2.68
14	SWH-05-20101106_16k-emission_swahi	ninaridhishwa na kila linaloendelea sasa kwani mambo ni shw	SWH-05-20101106/SWH-05-20101106_16k-emission_swa	16000	4.91
15	SWH-05-20101106_16k-emission_swahi	changamoto inayo tukabili kwa sasa	SWH-05-20101106/SWH-05-20101106_16k-emission_swa	16000	2.34
16	SWH-05-20101106_16k-emission_swahi	ni kutangaza matokeo ya uchaguzi huu	SWH-05-20101106/SWH-05-20101106_16k-emission_swa	16000	2.27
17	SWH-05-20101106_16k-emission_swahi	inabidi zoezi hilo lifanyike kwa amani pia	SWH-05-20101106/SWH-05-20101106_16k-emission_swa	16000	3.29
18	SWH-05-20101106_16k-emission_swahi	nimewaomba viongozi mbalimbali wa dini	SWH-05-20101106/SWH-05-20101106_16k-emission_swa	16000	2.39
19	SWH-05-20101106_16k-emission_swahi		SWH-05-20101106/SWH-05-20101106_16k-emission_swa	16000	3.4
20	SWH-05-20101106_16k-emission_swahi	jeje madii ni msemaji wa rais aliyeko madarakani lauren bagb	SWH-05-20101106/SWH-05-20101106_16k-emission_swa	16000	5.1

Figure 6; Cleaned and processed text file and data set.

After the preprocessing phase we saved the data in a new .csv file named Metadata_TZ.csv.

4.2. Future Work.

i. We are expecting to Train model after cleaning of data and transforming them into specific format then splitting them for Easy and accuracy model operation.

ii. We are expecting also to deploy our model into web-based application using flask framework.

iii. We are expecting to test and receive feedback for more modification of our model and Accuracy performance of our model.

References

Chen, Y.-H. a. (2019). Transcribear-Introducing a secure online transcription and annotation tool. *Digital Scholarship in the Humanities*.

Gelas, H. a. (2011). Evaluation of crowdsourcing transcriptions for African languages. *Laboratoire Dynamique Du Langage*.

Hadrien Gelas, L. B. (2012, 1 1). DEVELOPMENTS OF SWAHILI RESOURCES FOR AN AUTOMATIC SPEECH RECOGNITION SYSTEM. *Proceedings of the Joint Conference JEP-TALN-RECITAL*, pp. 1-2.

Himmelmann, N. P. (2018). Meeting the transcription challenge. *Reflections on Language Documentation 20 Years after Himmelmann*, 2-4.

Hyryn, O. (2020). Basic challenges in natural language processing systems. *Studia Philologica*, 41-45.

Ma, P. M. (2022). Natural Language Processing and Artificial Intelligence for enterprise management in the Era of industry 4.0. *Applied Sciences*, 1-2.

Masua, B. a. (2020). Enhancing text pre-processing for Swahili language: Datasets for common Swahili stop-words, slangs and typos with equivalent proper words. *Data in Brief*, 106-517.

Ranchal, R. a. (2013). Using Speech Recognition for Real-Time Captioning and Lecture Transcription in the Classroom. *Learning Technologies, IEEE Transactions on*, 299-311.

Robert Fay, W. T. (2021). Artificial Intelligence, Big data, security. *The Cyber Security Battlefield*, 1-8.