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.

1. 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.

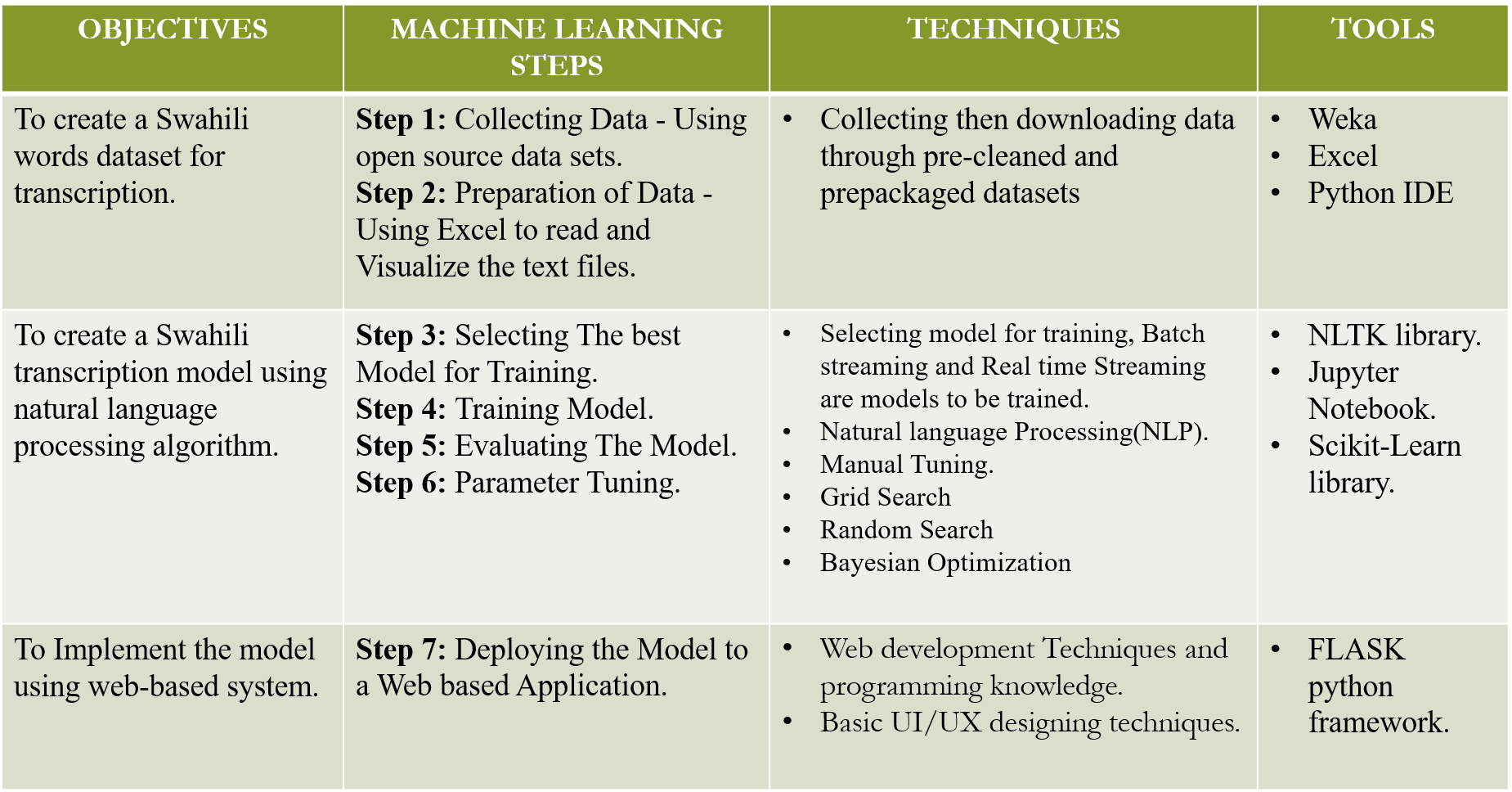
**Model Selection**: Select an NLP algorithm that will be used to transcribe the audio data. Some popular NLP algorithms for transcription include hidden Markov models (HMMs), recurrent neural networks (RNNs), and long short-term memory (LSTM) networks.

**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.

1. 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.



Project Methodology.