**Mental Health Chat bot documentation**

**Business Understanding**

**1. The client /user and their needs described**

The clients are in need of a virtual assistant using machine learning techniques with the aim of delivering mental support to kenyans. The chatbot is expected to have the capability to comprehend and evaluate the user's messages, and offer appropriate responses and resources, such as information on mental health conditions, methods of coping, and professional assistance.

**2. The client engagement process is provided**

This was done through clear reading of client demands on the links provided to us by the ML hackathon instructor: <https://arxiv.org/abs/1812.11783>

**3. The objective of the task**

We were tasked with building a chatbot utilizing machine learning techniques to furnish mental health assistance to individuals in Kenya. The chatbot should have the ability to recognize and assess user messages and provide suitable replies and resources, such as details on mental health conditions, coping strategies, and expert assistance.

**Data Acquisition**

**4. Sources clearly described**

We used the mental health conversational data downloaded from kaggle, to test and train our chatbot.

**5. Data acquisition process clearly described**

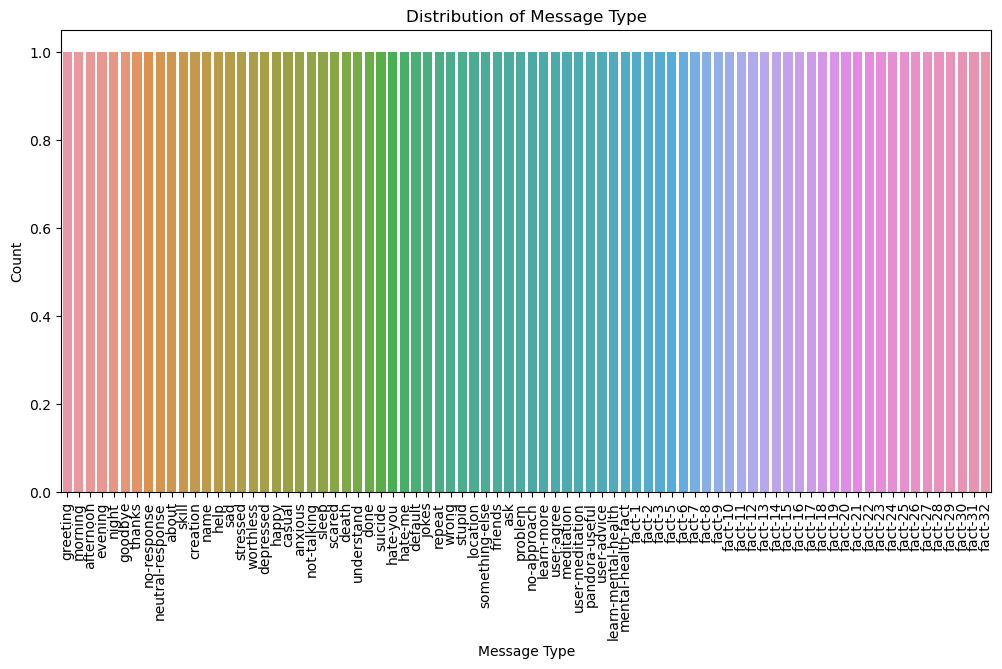
**Exploratory Data Analysis**

**6.0 The exploratory data analysis and outcomes**

EDA is the process of analyzing and summarizing the main characteristics of a dataset.

The main variable of interest is the message content.

The EDA techniques used included exploratory visualizations, to analyze the trends of the frequently asked questions over the period



Identifying patterns – This involves identifying related words and phrases and analyzing some of the relationship between different responses and questions.

### Feature selection – This involved selecting important features that was used to train the chatbot. This was involved selecting some of the most frequent phrases and word to train our model.

### Visualizations – This includes use of histograms and box plots to examine the distribution of message type across different intents.

DATA CLEANING

We used intents which involves categorizing the data into different intents or categories and then performing cleaning operations on each intent separately.

### DATA CLEANING PROCESS

* Categorized the data into different intents
* Created a JSON file and each intent had a tag, pattern and response
* Identified duplicated patterns and response and removed unnecessary characters and formatted the data
* We went through the entire datasets available line by line highlighting and extracting key information related to chatbot such as words and questions and their respective answers.

### DATA CLEANING OUTCOMES

* We were able to acquire a dataset that only contained chatbot related issues only. The dataset was then used to create a JSON file which was used to solve the problem.

*Figure 3: an example of JSON file*

## 

## **FEATURE ENGINEERING**

Feature engineering is the process of selecting and transforming raw data into features that can be used to train a machine learning model.

### FEATURE ENGINEERING PROCESS

After doing the data cleaning process and transforming data in a format of inbound and outbound text. The inbound text contained questions and queries from the users while the outbound texts contained responses from the support center to the users. This enabled us to remain with questions the users are most likely to ask in our training set.

These steps were followed for the feature engineering process

1. Converting data to lowercase

Converting data to lowercase is one of the useful preprocessing steps before training a chatbot. This ensures consistency by ensuring that all training data was represented in a consistent way. Consistency helps the chatbot to better recognize and words and phrases by reducing the number of errors that may occur due to case sensitivity. Converting data to lowercase also reduces the dimensionality of the data and ensures that the chatbot predicts the output with high speed.

1. Tokenizing of data

Tokenization is the process of breaking text into smaller units called tokes. We broke down the words and phrases into smaller units this transformed the data into a structured manner. For tokenizing our data, we used a word tokenization which involves breaking words into text. By breaking words into individual tokens, the algorithm was able to analyze text more effectively by identifying more frequent words and even helping the model predict the next word with high accuracy.