**CPIMS VIRTUAL ASSISSTANT**

**ALOYS ABOGE JR**

**ORIEL KIPLANGAT**

**GILKS MOSETI OMWOYO**

**OTIENO BILDAD MOSES OKOTH**

**Business Understanding**

The client in a CPIMS user support virtual assistant project is likely an organization working with children and using CPIMS. The client needs an automated solution to handle repetitive user support requests, which are time-consuming for their service desk staff. The virtual assistant should be user-friendly, accessible, available 24/7, handle a variety of requests, and provide fast and accurate answers. Input and feedback from CPIMS users will be critical to ensure the virtual assistant meets the client's needs.

We identified client needs by gathering data from a WhatsApp chat that contained inputs and feedback from CPIMS users. The inputs had similar questions from CPIMS users making the help desk repeat the same answers overtime. We ensured the virtual assistant met the client needs by getting him to aid in answering the common questions.

**Data Acquisition**

Data was collected from exported WhatsApp chats with Nairobi County CPIMS, Msa CCI’s & DCS CPMIS with CPIMS WAJIR TEAM, INSTITUTIONS CPIMS GROUP and BUNGOMA CPIMS GROUP

The data acquisition was quite simple as the dataset was readily provided by the CPIMS team.

**Explanatory Data Analysis**

The cleaned and prepared data was analyzed to identify

* patterns
* Trends
* insights

that can help inform the development of the virtual assistant.

This analysis could include techniques such as:

* data visualization
* Clustering
* natural language processing (NLP)

to identify common themes and patterns in user queries.

**Data cleaning and preparation**

The data was cleaned and prepared for analysis to ensure that it is consistent, accurate, and complete. It involved

* removing duplicates
* fixing errors
* standardizing data formats.

We used a cleaning tool called Pushbullet which converted the WhatsApp txt file to excel format.

We removed text messages with empty cells and irrelevant data

e.g. <Media omitted>, “This message was deleted” which were repeated multiple times.

We then read the 4,321 rows of data as we filtered the most frequent questions and then created new data sets of the questions with their respective answers. This provided us with a clear view of the question that was repeatedly asked and the ones which weren’t.

Afterwards we converted the excel folder to Json which we used for the subsequent processes e.g model training.

**Feature Engineering**

We employed discretization as we took the messy data that we had, which was pure text and emojis, emoticons alike and created two separate features possessing similar characteristics i.e. queries and responses which we again filtered based on what was relevant to us. We transformed our raw data into features that were used to build a machine learning mode in the following steps:

I. Data selection- we selected the most relevant questions from our cleaned dataset. The messages included queries and possible responses.

II. Feature Extraction- We extracted the queries from the dataset depending on the nature of the queries. We then transformed the extracted queries into Json format that could be used to train the machine learning model.

III. Feature Engineering- Based on the insights gained from data exploration and analysis, we created a data model for the virtual assistant that includes a list of common user queries and the appropriate responses. This model would be designed using machine learning (ML) and NLP algorithms to enable the virtual assistant to recognize and respond to user queries accurately and efficiently.

III. Feature Selection- The queries were selected and grouped based on how they related with each other. We finally had categories such as greetings, resetting passwords etc.

IV. Training the Model- The data model was trained and tested using historical data and simulated user queries to ensure that it accurately recognizes and responds to common user queries. The model would need to be continually refined and improved based on feedback from user interactions with the virtual assistant.

V. Model Evaluation- After training the model using the labeled dataset , it achieved an accuracy of 94.5%, which indicates that it is able to correctly classify support requests and problems for the users

VI. Model Deployment-

**Model Development**

Our task is centered around the imitations of the workings of a neural network. We use both python and tensorflow majorly the keras, a high-level library running on top of tensorflow.

It is a pretrained model as we fed it data from the corpus we had previously cleaned. We performed the following actions on the chatbot to attain optimum efficiency:

1. Importing the required libraries .ie. tensorflow, numpy, pandas,json,nltk.

2. Importation of the cleaned data which was in now in a json format

3. Pre-processing the data: employing use of e.g. tokenization, pad\_sequence.

4. We take the data, look at the number of unique words

5. We pass the data as numbers across the embedded layer and output a 2d vector (word with its corresponding vector).

6. With an epoch = 200, we manage to get an accuracy of .94 i.e. roughly 95%.

7.We test our model by, in our case, asking the FAQ’s.

**Model Evaluation**

Our model passed with an accuracy of .94 with the accompaniment of epoch = 200. We tested by asking the questions the data revealed were the most frequent. The answers were more than fairly accurate although not entirely. We experienced challenges of which most we addressed e.g. the mixture of different languages i.e. Swahili and English.

**Partial Solution:** We employed the use of keywords in an attempt to nullify whatever challenge the change in language posed.

**Observation**: We would have to enrich our corpus in an attempt to increase the bot’s knowledgebase and familiarity with the prospective questions.

**Model Deployment**

We implemented our model as a CPIMS virtual assistant alias Maya. She is the product of the model we created and will hopefully pass all the required tests to be an active virtual assistant to the CPIMS organization.

**Challenges Faced in Handling the Task**

1. A lot of the data provided was similar in terms of the questions asked and so was the answers given e.g. resetting the password. This greatly reduced the variability of questions and responses we could feed our chatbot.
2. The data we were provided with proved to be real difficult to work with as the format (exported WhatsApp chat) needed a lot of time to clean.
3. Conversion of the file format to suitable format e.g. txt to xlsx to csv to json.
4. Transferring the output of the bot to the web needs alot of time and proper implementation.

**REFERENCES**

Subramanian, D. (2019, December 16). Synonyms and antonyms in Python.

Saleem, A. (2022, August 16). Building a rule-based chatbot in Python. Data Science