



# Destina Automatic Speech Recognition Conversational AI Assistant

26.04.2022

Jacques Patricks Khisa

Savvy Technologies

Kilimani, 8th Floor, Pinetree Plaza

Nairobi.

## Overview

This is a documentation for deploying a production-level conversational AI to enable customer service in retail services for small and medium-sized businesses using automatic speech recognition. It also seeks to establish explainable artificial intelligence in the natural language process tasks as it applies large transformer models and a multilingual domain dataset for retail of English, Kiswahili and Kikuyu.

## Goals

1. Curate a rich contextual multilingual dataset for the retail services domain.
2. Deploy and enable pretrained Automatic Speech Recognition and Named-Entity Recognition models on NVIDIA RIVA.
3. Deploy a production-level conversational AI application with Helm Chart for scaling on Kubernetes.
4. Integrate communication channels and continuous deployment.

## Specifications

Building conversational AI will have both functional and non-functional requirements.

### Functional Requirements

API Key from NVIDIA NGC, Language models from HuggingFace and NVIDIA Train Adapt and Optimize Kit for transfer learning, NVIDIA NeMo for a named-entity recognition, NVIDIA RIVA for deployment and NVIDIA Triton for inferencing.

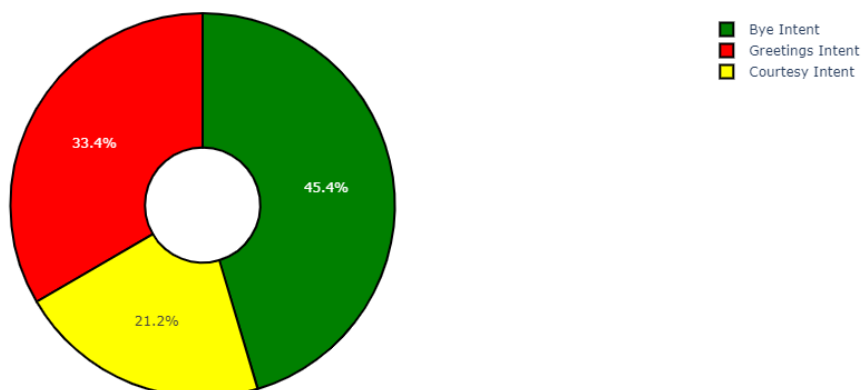
## Milestones

### I. Intent Predictions on Curated Sheng Corpus using Facebook Bart-Large-MNLI HuggingFace Language Model.

I collected raw textual data from social platforms like WhatsApp and Twitter to simulate real conversations and classified them into intents. Using zero-shot classification as my NLP task, I supplied the `facebook-bart-large-mnli` HuggingFace language model with the corpus to obtain prediction performances on the intent labels. The results were as follows:

1. Performance on common intents (Greetings, Courtesy, Bye).

Chitchat intentions: Predictions With Facebook Bart-Large-MNLI Model on Sheng Curated NLU

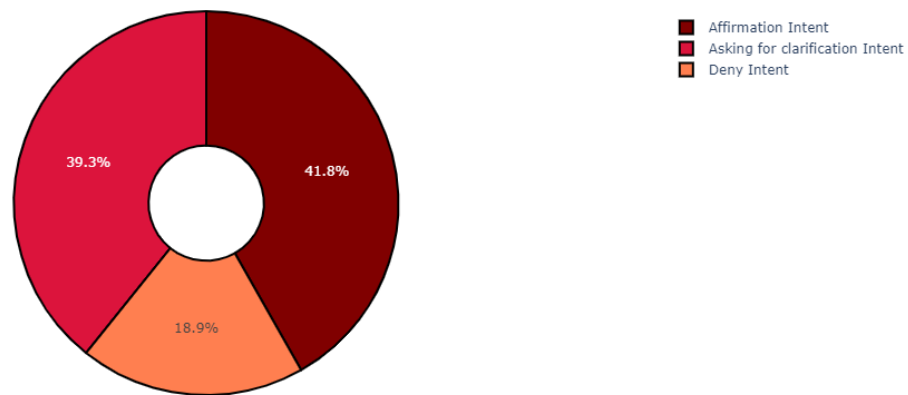


The model performs below par on predicting the multilingual intents. Even so, it performs higher on **Bye** intent than on **Greetings** and **Courtesy**. Performs lowest in predicting **Courtesy** intent.



## 2. Performance on conjunction (cause dialog turns) intents.

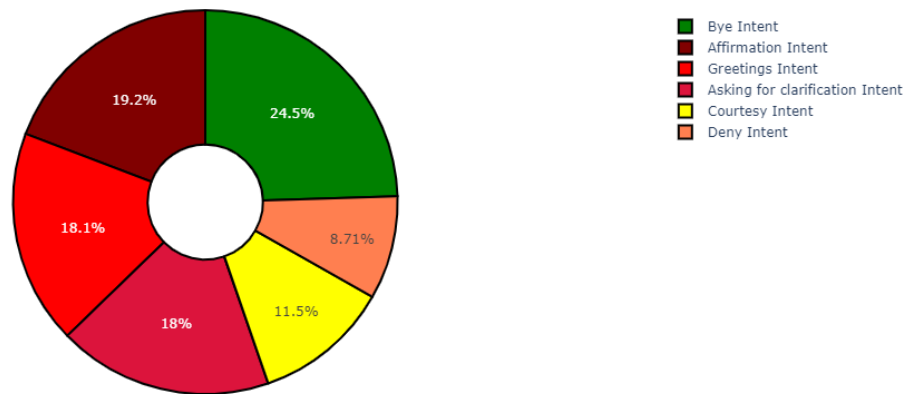
Affirm, Deny & Clarification intentions: Predictions With Facebook Bart-Large-MNLI Model on Sheng Curated NLU



The language model generally performs below average on all intents with higher performance on predicting **Affirmation** intent in contrast to **Asking for clarification** and **Deny** intents. Performs lowest on predicting deny intent.

### 3. Performance on six top intents in a normal conversational flow.

ChitChat Intentions: Predictions With Facebook Bart-Large-MNLI Model on Sheng Curated NLU



The facebook bart-large-mnli language model has an average prediction score of 18%. This is significantly low.


## II. Intent Predictions on Curated Sheng Corpus using XML-RoBERTa-Large-XNLI HuggingFace Language Model.

This is a contrasting model that tries to mitigate the shortcomings of the Facebook Bart-Large-MNLI model. This model is intended to be used for zero-shot text classification, especially in languages other than English. It is fine-tuned on XNLI, which is a multilingual NLI dataset. The model can therefore be used with any of the languages in the XNLI corpus which includes Swahili and English. Since the base model was pre-trained on 100 different languages, the model has shown some effectiveness in languages beyond English.

1. Performance on common intents that drive normal dialog turns/ conversations.

Curated Sheng NLU Intent Predictions With HuggingFace: XLM-RoBERTa-LARGE-XNLI Mo

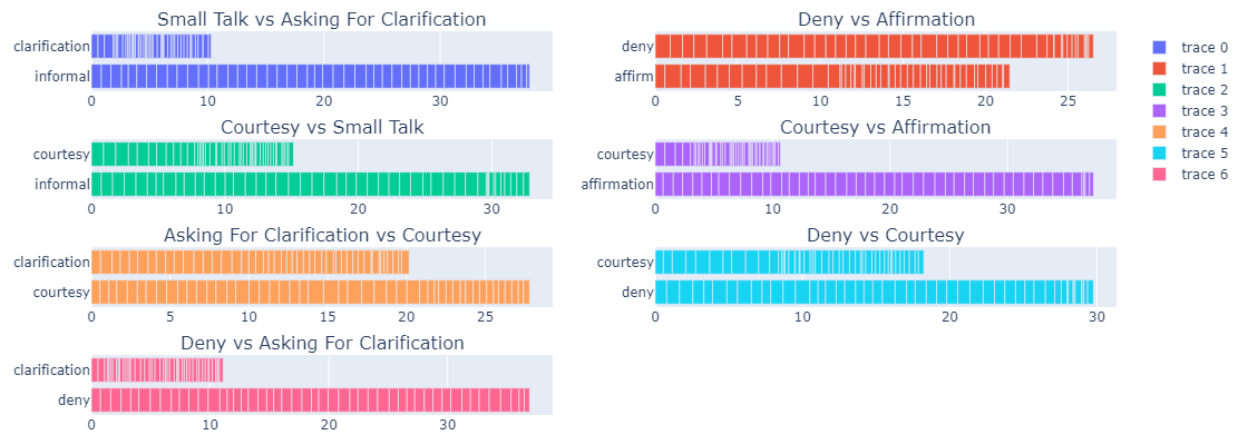




The language model predicts the **greetings, bye, deny, courtesy** and **affirmation** intents correctly but has trouble predicting the **asking for clarification** intent. Even so, the correct prediction on deny intent is slightly higher than the wrong prediction on affirm intent. This performance is also exhibited on the correct prediction of the affirm intent vs the wrong prediction on deny intent. This could be as a result of limited data provided to clearly isolate the intents to enable the model. Therefore, the model calls for more data in the dissimilar intents.



## 2. Performance on intent pair comparisons with increased size of the corpus.



The model performs much better with these comparisons in various ways. It predicts **small talk**, **deny** and **affirmation** intents correctly. The prediction on **courtesy** vs **small talk** intents and **asking for clarification** vs **courtesy** is incorrect. This could be a reason to merge courtesy and small talk as a compound intent. In contrast, asking for clarification and courtesy are polar ends. Therefore, the xlm-roberta-large-xnli is a better language model for sheng NLU predictions!