# FSM Online Internship Completion Report
on

# INTP23-ML-1: Chatbot for FSM
In

Machine Learning

Submitted by

Nikita Sharma
Jaypee Institute of Information Technology, Noida

Under Mentorship of
Keivalya Pandya

FSM
Smart Manufacturing Simplified

IITD-AIA Foundation for Smart Manufacturing

[1-June-2021 to 31-July-2021]

Chatbot for FSM

# Chatbot for FSM


## Abstract

Industry 4.0 represents a production concept based on automation, real- time optimization and digitalization of production factories. Its main driver is Artificial Intelligence (AI) and Machine Learning (ML); capable of handling large dataset and identifying human patterns in everyday life. In this context, Generative Pre-Trained Transformer (GPT), Assistants have great impact and importance in Industry 4.0 for Natural Language Processing (NLP), code generation, pattern identification. The end-user may require assistance for operating procedures, equipment operations and resolving common errors, for example Frequently Asked Questions (FAQs) to rectify errors or to start the process and this can be done through providing them with assistance and real-time queries resolving. Real time monitoring, GPTs can manage organizations' data or information for data collection, monitoring and output efficiency. The main objective is we have focused on designing a GPT that will assist queries related to IAFSM by scrapping and learning all organizations' data/information using sources like organizations' website. However, GPTs can use data like previous chat logs, service log etc. which will provide aid to users, based on more training data. GPT will assist the user queries; providing them with documentation, video links. These queries/ resource can be linked with data i.e., conditions, outputs, inputs value and storing them for future references and predictions related to defects and maintenance by providing alerts, messages through messenger Applications (Apps) or web. This method will enhance user experience, provide 24X7 support.



Keywords: Deep Learning, Natural Language Processing, Industry 4.0, Artificial Intelligence, Chatbot, Chat GPT, Transformers

# Table of Content

## Introduction to the problem

In the support service, one of the most important aspect is quality and precise supports. Customers go through a lot of questions and problems from technological to data safety, and lack of support at any stage may prevent user/customer from using that particular technology making it a failure. With the rise of technology every day, the demand for quality customer support, better user experience is rising. Due to higher number of issues, large number of queries remain unanswered sent through emails and call centers. This is where Machine Learning and Natural Language Processing can provide an alternative for Chat-support and user experience to support other service channels. By focusing on Words and sequence Analysis of queries and context, it can be identified "what is been asked by user" and quick service can be provided. Essentially, deploying it to the web for remote access can ensure accessibility to a larger group of industry people.

With the continuous development of Natural Language Processing, Transformers models have been used in context understanding for large quantities of data. To establish a better connection between customer and company, we propose a Virtual Assistant using Transformers "BERT" fine-tuned on site data.

## Literature Review

### Accuracy

- ✓ I have used two models, DistilBERT for predicting the answers and confidence score of BERT model to get the exact answers from all predicted ones.
- ✓ Also, string search selecting the answer having string or target _word (topic asked in question) in the lines.

### Generative models

Three types of models are famous and successful can be used:

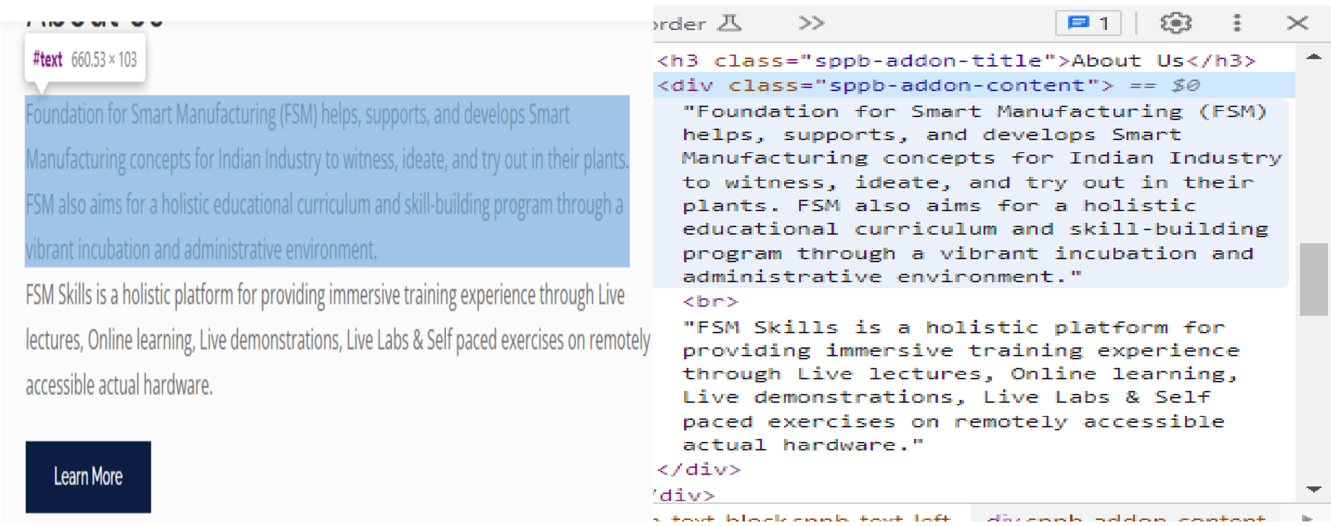
- LangChain model- LangChain is an open-source framework for developing applications powered by language models. Once you have a language model, you need to create a chain. A chain is a sequence of instructions that tells the language model what to do.
- Large language model- A large language model (LLM) is a language model characterized by its large size. LLMs are trained on massive datasets of text and code, which allows them to learn the patterns and connections between words. This allows them to perform a variety of tasks,
- Rule-Based- in this for each and every type of question e.g., greet, interrogation, etc., we have to define the answer manually.

We are using Large Language Models, in this project.

# Dataset Analysis

## About the dataset

Web scrapping was done using beautiful soup, and heading, paragraphs was extracted. some divisions directly have the data without any html element. As given in the fig (i) all these types of divisions have same class “sppb-addon-content” so divisions with this class name was searched and printed their content. Now, there are links which direct to other page for example button named “Learn more”. Figure (ii) explain How scrapping of links was done. Now the scrapped data was in raw format. So, all double inverted commas are changed with single inverted commas otherwise there is confusion about context start and end, also extra spaces, tabs (/t) and next lines (/n) symbols, repeated content was removed in order to reduce the size of context from 38 K words to 16 K and then to 11K as the CPU was taking too much time for processing. Context scrapped was in json format.




(fig i)

***⌊Dashboard (1st link)***

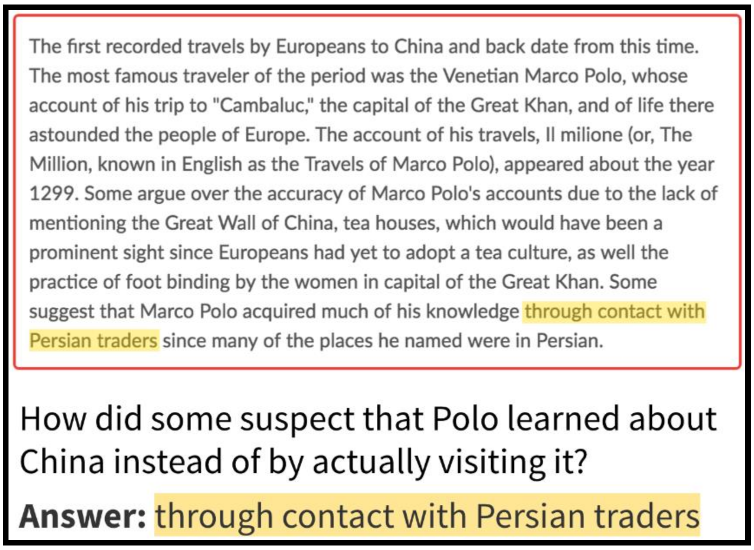
    ***⌊ all links on dashboard (eg. Links of Technologies, services, facilities)***

        ***⌊ all links on redirected Page (eg. All the links on Services page)***

(fig ii Link Extraction)

## SQuAD

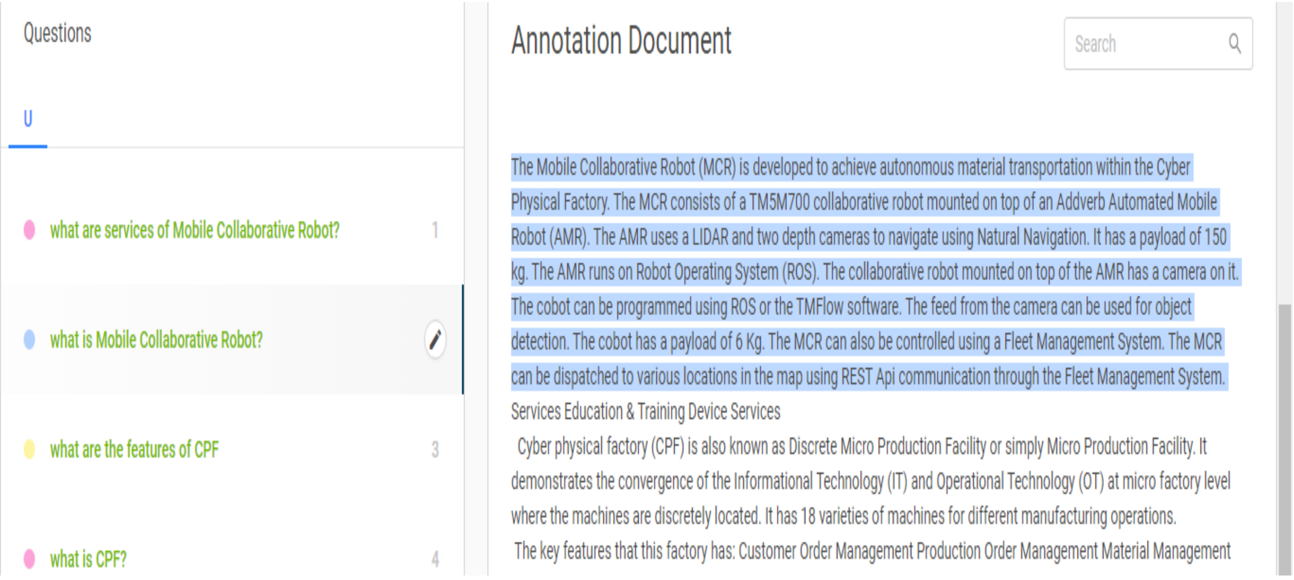
SQuAD stands for Stanford Question Answering Dataset. It is widely used in the NLP (Natural Language Processing) field for training and evaluating the NLP models used for question-answering. It has context, question and their corresponding answers. The goal is to accessing the models' capability of reading the comprehensions and extracting correct answer to questions from provided passage. Answers can be one word and some group of words.




(fig (iii) General example to explain about SQuAD)

### Data Preparation

In order to increase the accuracy of models fine tuning is performed. I used Haystack Annotation Tool for preparing the SQuAD format data of IAFSM website for training the models where for each context (512 token length) have to frame questions and mark its respective answer.




Fig(iv)Answer highlighted for the question what is mobile collaborative robot


Fig (v) website data in squad format

I used following code to extract the question, answer, context from the dataset

```
# importing the json file to read the data in order to finetune the model
import json
from pathlib import Path

def read_squad(path):
    path = Path(path)
    with open(path, 'rb') as f:
```

```
        squad_dict = json.load(f)

    contexts = []
    questions = []
    answers = []
    for group in squad_dict['data']:  # this is done according to Squad dataset
        for passage in group['paragraphs']:
            context = passage['context']
            for qa in passage['qas']:
                question = qa['question']
                for answer in qa['answers']:
                    contexts.append(context)
                    questions.append(question)
                    answers.append(answer)
    #print(questions[:5])
    return contexts, questions, answers

train_contexts, train_questions, train_answers =
read_squad('more_answers_squad_data.json')
```

### Deployment in hugging face

I used Gradio in order to develop the user interface for chatting or asking queries and deployed on spaces.

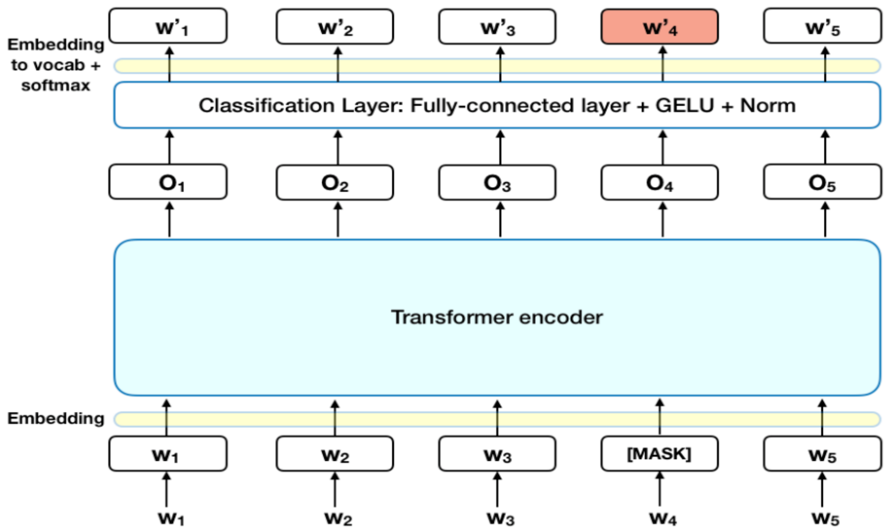
**Final deliverable app:**

LINK: https://huggingface.co/spaces/Nickitaa/gradio-chatbot

## Algorithm Explain

### Overview of Transformers:

Transformer is used when there is large context to understand. It consists of three things namely, encoder, decoder, Neural network. In this; question, context is provided to encoder. And there is neural network in between encoder and decoder. The weights of neural network provide the information which word in context need to be addressed and, has some importance, wrt context relation. These words (vector) are then passed to decoder which convert tokenized value to actual word and previous words is used to predict current word.




(fig vi)

**BERT** is a type of transformer-based model, which handles sequential data, such as text. Its bidirectional pretraining approach, which means that the model is trained on both left-to-right and right-to-left contexts of a given text enables BERT to capture a deeper understanding of the relationships between words in a sentence. It is pre-trained to predict missing words in sentences and the next sentence in a pair. After pretraining, it can be fine-tuned to train on specific data. Source: bert-large-uncased-whole-word-masking-finetuned-squad

**DistilBERT**, a distilled version of BERT: smaller, faster, cheaper and lighter. DistilBERT is a small, fast, cheap and light Transformer model based on the BERT architecture. Knowledge distillation is performed during the pre-training phase to reduce the size of a BERT model by 40%.
Source: twmkn9/distilbert-base-uncased-squad2

## System and Hyperparameter detail

Our operating system is windows 8 Pro, using the Pytorch framework, and training and testing on CUDA, CPU.

- Learning rate=1e-4
- Batch size=16
- Number of epochs = 3
- Number of iterations=100
- Max Length=512
- Vocab size=30522
- Activation =gelu
- Model-type=distil-bert
- Stride=100

## About the dataset

SQuAD format is used in training the model. My model is fine-tuned on three length of answer.

1. Stanford Question Answering Dataset which has one word or some group of words as answers, model is pre-trained on this dataset.
2. Website information in the squad format is saved in json file that is used to fine-tune the model. Which has length of 100 words e.g. complete process of internship, etc.
3. In order to increase the accuracy of model third data is prepared which has the answers length of 20-30 words.

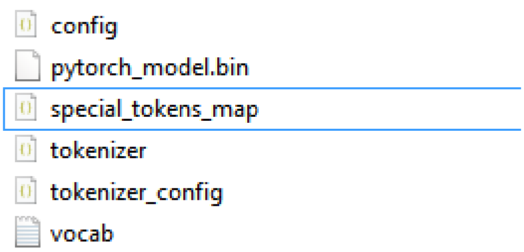
On combining whole fine -tuning data 160+ questions were there from site itself.

## Procedure followed for fine-tuning

Web-scrapping
|_ extracted json file
|_pasted the data to word to convert into correct context(cleaning)
|_used context for SQuAD format data generation
|_Trained the model and saved it
|_ Loaded the saved model and again trained it

## Model Structure

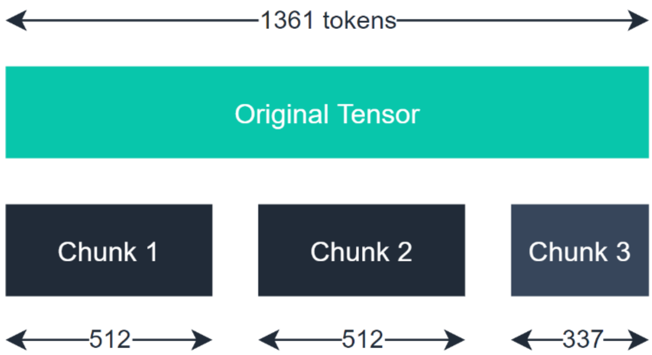
model
|_distil-bert custom
|_



## Innovation in Implementation

1. I used fine-tuned DistilBERT model for predicting answers as fine-tuning was done using CPU, so it has less parameter and fast in implementation and BERT for Confidence score for choosing the correct answers from all the predicted answers of DistilBERT.

2. **2**.BERT model can only read 512 tokens (including two special tokens) at a time and truncate rest of the context. Divided context of 15K+ words in BERT model into chunks each of 512. Fig(vii) However last chunk can have less than 512 tokens. Stride=100, to repeat the last 100 words of previous context in current context.

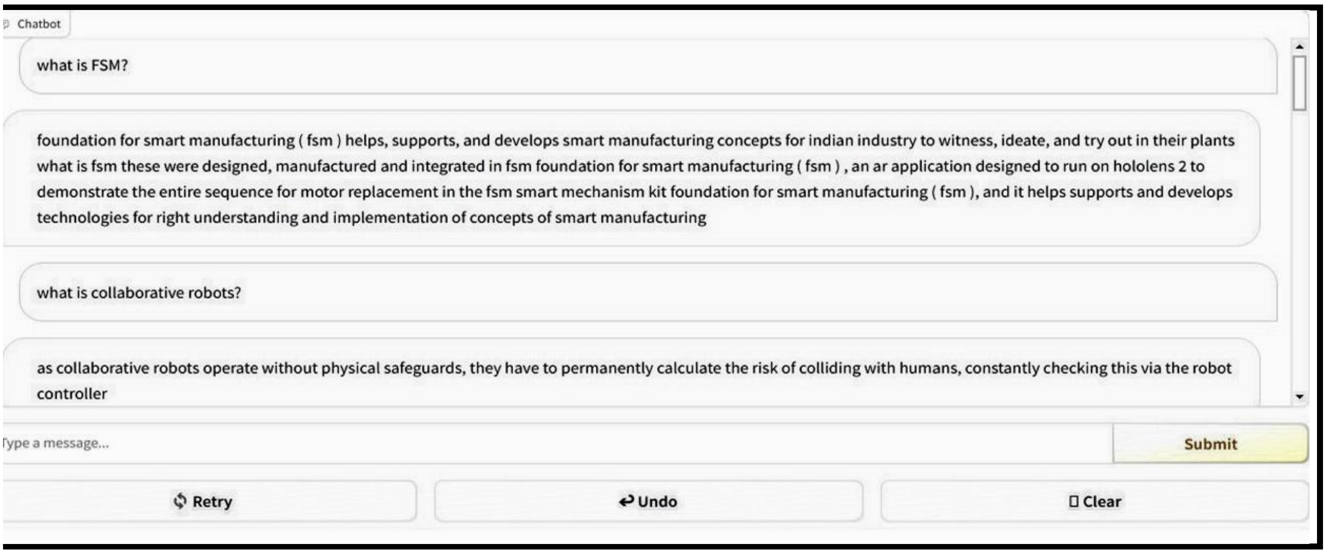


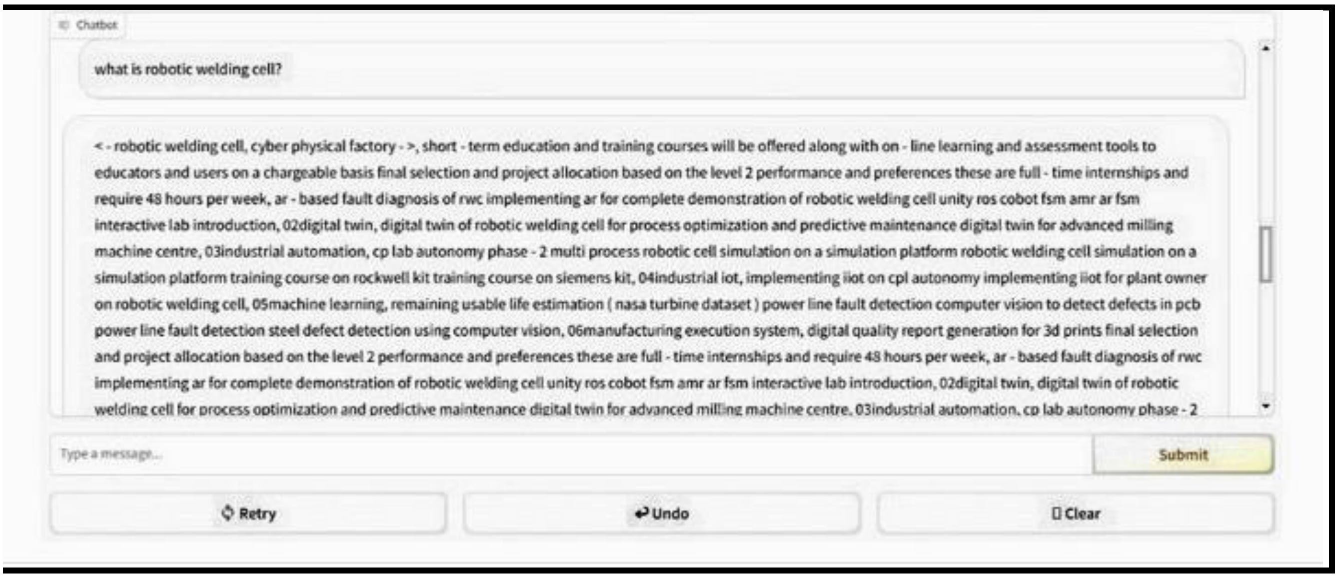

Fig(vii)

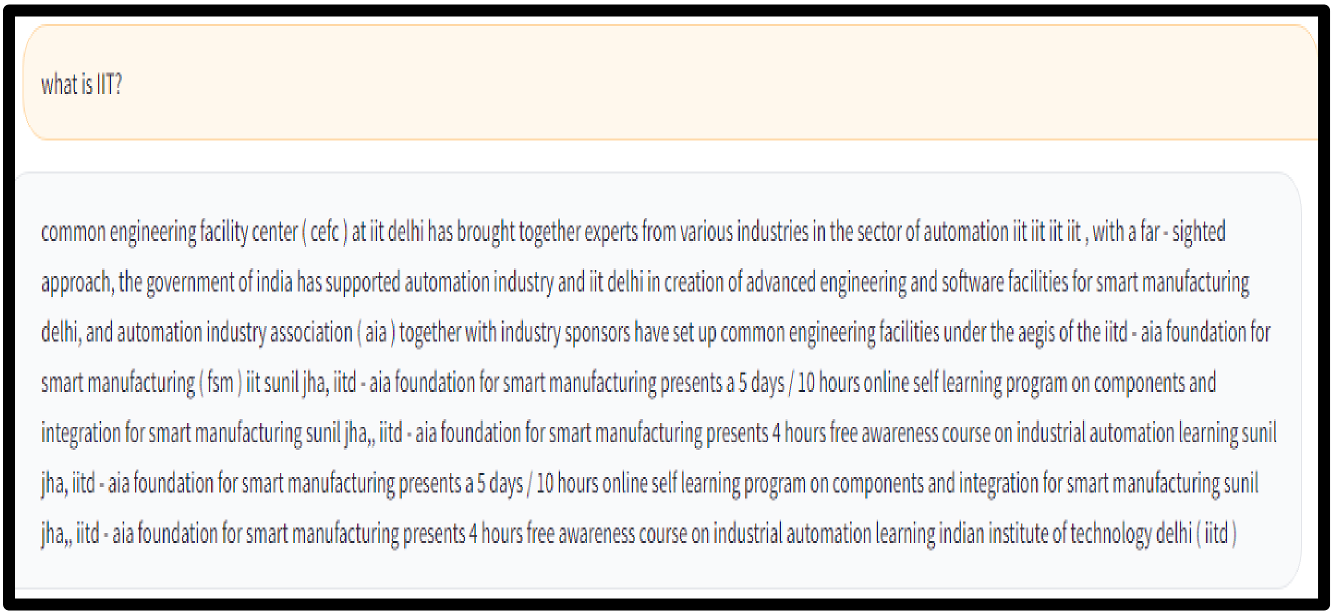
3. After fine-tuning and training the model, bot was giving answers but there was some extra stuff for example Que: “What is fsm?” and the answers were with information about robots, welding cell etc.
   So, in order to solve this, I used two things:
   1) Confidence score
   2) String search

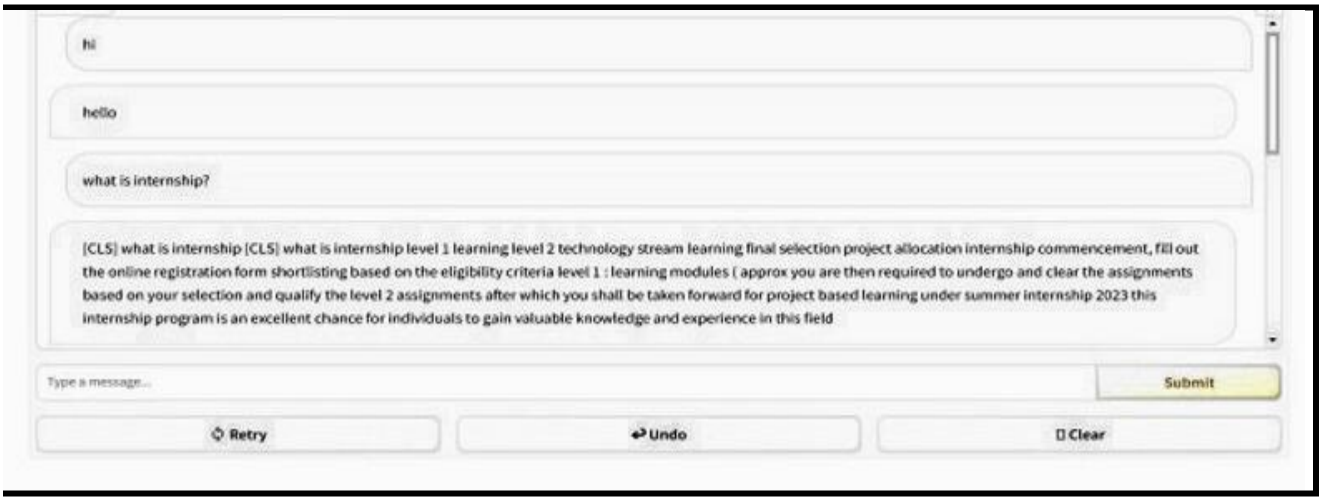
I used a pipeline Confidence score to search for answer that has correct context with respect to question i.e., answer with high score.
But within that answer there were some extra topics e.g. “With information about internship. Next line is about collaborative robot.” Now suppose question was about collaborative robot so we have to extract that line having collaborative robot context for that I extracted out target_word from question. In order to get that target_word around which question revolves for example “what is fsm?” so “fsm” is target_word. I tokenized question and now made a list of “helping verbs, wh-family words” etc., and the words not in list are considered as target_word and line having these word are part of the exact final answer.

# Results


Chatbot
what is FSM?
foundation for smart manufacturing ( fsm ) helps, supports, and develops smart manufacturing concepts for indian industry to witness, ideate, and try out in their plants what is fsm these were designed, manufactured and integrated in fsm foundation for smart manufacturing ( fsm ) , an ar application designed to run on hololens 2 to demonstrate the entire sequence for motor replacement in the fsm smart mechanism kit foundation for smart manufacturing ( fsm ), and it helps supports and develops technologies for right understanding and implementation of concepts of smart manufacturing
what is collaborative robots?
as collaborative robots operate without physical safeguards, they have to permanently calculate the risk of colliding with humans, constantly checking this via the robot controller
Type a message...
Submit
Retry
Undo
Clear


Chatbot
what is robotic welding cell?
< - robotic welding cell, cyber physical factory - >, short - term education and training courses will be offered along with on - line learning and assessment tools to educators and users on a chargeable basis final selection and project allocation based on the level 2 performance and preferences these are full - time internships and require 48 hours per week, ar - based fault diagnosis of rwc implementing ar for complete demonstration of robotic welding cell unity ros cobot fsm amr ar fsm interactive lab introduction, 02digital twin, digital twin of robotic welding cell for process optimization and predictive maintenance digital twin for advanced milling machine centre, 03industrial automation, cp lab autonomy phase - 2 multi process robotic cell simulation on a simulation platform robotic welding cell simulation on a simulation platform training course on rockwell kit training course on siemens kit, 04industrial iot, implementing iiot on cpl autonomy implementing iiot for plant owner on robotic welding cell, 05machine learning, remaining usable life estimation ( nasa turbine dataset ) power line fault detection computer vision to detect defects in pcb power line fault detection steel defect detection using computer vision, 06manufacturing execution system, digital quality report generation for 3d prints final selection and project allocation based on the level 2 performance and preferences these are full - time internships and require 48 hours per week, ar - based fault diagnosis of rwc implementing ar for complete demonstration of robotic welding cell unity ros cobot fsm amr ar fsm interactive lab introduction, 02digital twin, digital twin of robotic welding cell for process optimization and predictive maintenance digital twin for advanced milling machine centre, 03industrial automation, cp lab autonomy phase - 2
Type a message...
Submit
Retry
Undo
Clear


what is IIT?
common engineering facility center ( cefc ) at iit delhi has brought together experts from various industries in the sector of automation iit iit iit iit , with a far - sighted approach, the government of india has supported automation industry and iit delhi in creation of advanced engineering and software facilities for smart manufacturing delhi, and automation industry association ( aia ) together with industry sponsors have set up common engineering facilities under the aegis of the iitd - aia foundation for smart manufacturing ( fsm ) iit sunil jha, iitd - aia foundation for smart manufacturing presents a 5 days / 10 hours online self learning program on components and integration for smart manufacturing sunil jha,, iitd - aia foundation for smart manufacturing presents 4 hours free awareness course on industrial automation learning sunil jha, iitd - aia foundation for smart manufacturing presents a 5 days / 10 hours online self learning program on components and integration for smart manufacturing sunil jha,, iitd - aia foundation for smart manufacturing presents 4 hours free awareness course on industrial automation learning indian institute of technology delhi ( iitd )


hi
hello
what is internship?
[CLS] what is internship [CLS] what is internship level 1 learning level 2 technology stream learning final selection project allocation internship commencement, fill out the online registration form shortlisting based on the eligibility criteria level 1 : learning modules ( approx you are then required to undergo and clear the assignments based on your selection and qualify the level 2 assignments after which you shall be taken forward for project based learning under summer internship 2023 this internship program is an excellent chance for individuals to gain valuable knowledge and experience in this field
Type a message...
Submit
Retry
Undo
Clear

## Scalability to Solve Industrial Problem

In the fourth industrial revolution, Smart and Samarth technologies are evolving and NLP will provide a hand for development.

- In transformer model one of the important step for behavior of model towards prediction is fine tuning, model can be fine-tuned with customized data to suit specific requirements of industry and it can be domain specific and more relevant to industry.
- Transformers model pre-trained in various international language are available in the industry. So, it can be deployed round the globe and international collaboration will be easy.
- It can be relevant in assistance and support service, which the model is doing now i.e., aiding the user queries but it can aid developers in particular domain with respect to data provided.
- If provided with resources like cloud, better processing units it can handle and can be accessible to wider audience with less execution time

## Future Work

- I used Gradio as user interface but in order to increase the scalability of the project we can use WhatsApp as notification app.
- Model is fine-tuned specifically on website data, by using different data like project under IAFSM, research paper will increase the diversity and also the accuracy of the model.

## Conclusion

Chatbot using BERT model is able to answers queries of user. Using Gradio we can generate easy and Fast User interface and on hugging Spaces it can be deployed.

## References


- https://huggingface.co/
- https://www.analyticsvidhya.com/
- https://medium.com/
- https://annotate.deepset.ai/index.html