International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 09 | Sep 2023 www.irjet.net p-ISSN: 2395-0072

Revolutionizing Industry 4.0: GPT-Enabled Real-Time Support

Nikita Sharma[1], Keivalya Pandya[2]

[1]Jaypee Institute of Information Technology, Noida, India
[2]Birla Vishvakarma Mahavidyalaya Engineering College, Gujarat, India

---

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 datasets and identifying human patterns in everyday life. In this context, Generative Pre-Trained Transformer (GPT), Assistants have a significant impact and importance in Industry 4.0 for Natural Language Processing (NLP), code generation, and 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. This can be done by 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 training data. GPT will assist the users' queries; providing them with documentation, and video links. This is to design a GPT that will assist queries related to IAFSM by scrapping and learning all organizations' data/information using sources like organizations' websites. However, GPTs can use data like previous chat logs, service logs, etc. which will provide assistance to users, based on more queries/ resources that can be linked with data i.e., conditions, outputs, and input value, and stored for future references and predictions related to defects and maintenance by providing alerts, and messages through messenger Applications (Apps) or the web. This method will enhance user experience, and provide 24x7 support.*

Key Words: **Natural Language Processing, Chatbot, Transformer, Industry 4.0, Question Answering, Deep Learning**

1.INTRODUCTION

Customer service is essential in any industry. Due to the limitations of the companies for handling a large number of users/customers queries, we need some support system for 24x7 hours. In recent years, chatbots in the industry have maintained premium services. Chatbots with specific guidelines can only answer particular types of queries, not all real-time queries. However, with the concept of Deep Learning and Natural Language Programming (NLP) chatbots' ability to answer like humans will increase. Natural Language Processing (NLP), Natural Language Understanding (NLU), Natural Language Generation (NLG), etc., will boost the popularity of these chatbots for their finest support to the users. Deep Learning can be used to train the chatbot for different business and industrial domains.

In 2017, Transformer was introduced by the Google Brain team, and in 2018 BERT (Bidirectional Encoder Representations from Transformers) by researchers at Google led the chatbots to better understand context, multi-lingual, less labelled data for precise and more accurate answers generation. This paper explores the use of NLP in virtual assistants for answering user queries.

2. RELATED WORK

NLP is now the most researched area like sarcastic comments and chatbots as it improves human and computer communication. Chatbot applications are mainly in the fields of health, industries, entertainment, research, etc. Akhtar et al analyzed the chat conversations between customers and the chatbot of a telecommunication company to find out if these interactions can be used to determine a) users' topics of interest and b) user satisfaction [1]. To reach this goal, chat conversations are interpreted as sequences of events, and user inputs are analyzed with the help of text mining techniques. There are four different types of chatbot applications namely, service, commercial, entertainment, and advisory chatbot. Chatbots can be used in task-oriented and non-task-oriented [2]. Virtual assistants like Siri and Alexa are task-oriented and can perform tasks like making a phone call, playing a song, etc. Companies' websites have chatbots for answering user queries, sometimes based on certain specific technology. Non-task-oriented chatbots that tell the weather information can talk like friends and are used for entertainment purposes, for example, Replika: My AI Friend, Eviebot by Existor. Eviebot according to creators is the most popular artificial personality having perfect timings of facial expression and movement. Open-domain is when the chatbot or assistant can answer queries of users in a large number of fields, for example, Meena, Bard by Google, ChatGPT by Open AI, etc. Adiwardana et al [3] introduces Meena a multi-turn open-domain chatbot trained end-to-end on data mined and filtered from public domain social media conversations that can chat like humans. Closed domain [4] is when the chatbot can answer queries in a particular field related to health, company website, and education and have comparatively better results. Nakano et al [4] presented HRIChat- a framework for closed-domain chat that can be combined with a task-oriented dialogue system for better uses Long-short-term memory (LSTM) [5] is also a technique that can be used for language models used to reduce

traditional RNN problems. However, it has a range of memory and deep LSTM is introduced having multiple layers; used to process output from the previous layer as input and increase in terms of performance and can learn more abstract features. Sutskever et al [5] used a multilayered LSTM to map the input sequence to a vector of a fixed dimensionality, and then another deep LSTM to decode the target sequence from the vector. They aim to translate English into French. Liu et al [6] presented a novel framework that combines a bidirectional long short-term memory (Bi-LSTM) network and transformer to solve the problem, where the self-attention mechanism is substituted with Bi-LSTM to capture the semantic information from sentences. Meanwhile, an attention mechanism model is applied to focus on those important words and adjust their weights to solve the problem of long-distance information loss. Sung et al [7] presented a paper, that explored ways of improving the pre-trained contextual representations for the task of automatic short answer grading, a critical component of intelligent tutoring systems. Also, the authors provide Empirical evaluation on multi-domain datasets showing that task-specific fine-tuning on the pre-trained language model achieves superior performance for short answer grading. In 2017, Viswani et al [8] introduced the best-performing models connecting the encoder and decoder through an attention mechanism. They proposed a new simple network architecture, the Transformer, based on attention mechanisms. Experiments on two machine translation tasks showed these models are superior in quality while being more parallelizable and requiring less time to train. In 2018 BERT (Bidirectional Encoder Representations from Transformers) [9] was introduced by Devlin et al . According to the authors, BERT is designed to pre-train deep bidirectional representations from the unlabeled text by joint conditioning on both the left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

3. METHOD

Transformer is used when there is a significant context to understand. It consists of two things namely, an encoder, and a decoder. The encoder works on input sequences by converting them to tokens or sub-words There is a neural network between the encoder and the decoder. The weights of neural networks provide the information on which word in context needs to be addressed and have some importance concerning context and question relation. These words (vectors/tokens) are then passed to the decoder which converts the tokenized value to the actual word and previous decoded words are used to predict the current word. BERT is a transformer-based model that handles sequential data, such as text. Its bidirectional pre-training 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.[9] DistilBERT, is a smaller, faster, cheaper, and lighter transformer based on BERT architecture.

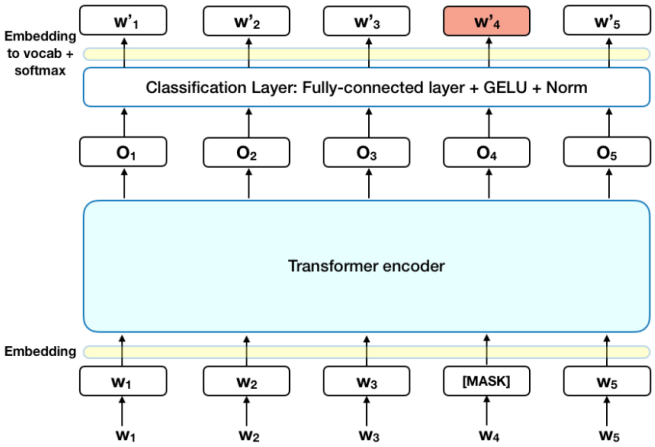

Fig -1: BERT Architecture

3.1 Data Collection

The IITD- AIA Foundation for Smart Manufacturing (IAFSM) site data is extracted through web scraping headings, paragraphs, and divisions. Links, redirect to another page, for example, the button named “Learn more” is scraped through Figure 2. All double quotation marks have been replaced with single quotation marks to remove the confusion about the context start and end point, extra spaces, tabs (/t), next lines (/n) symbols, and repeated content are removed from the data obtained through scraping to reduce the size of context from 38 K words to 11 K words. SQuAD is widely used in NLP for training and evaluating the models used for question-answering. The goal is to assess the models’ capability of reading comprehension and extracting correct answers to questions from provided passages. Answers can be one word or some group of words.

3.2 About the Dataset

Data Preparation: The model is fine-tuned to increase accuracy. I used the Haystack Annotation Tool for preparing the data in SQuAD format for training the models where in each context (512 token length or less) we have framed questions and marked their respective answers. “Figure 6” below shows the “Annotation Document” which is the context provided, a part of length 512 words or less from complete data of length of thousand words. The highlighted part is the answers to the question “What is a Mobile Collaborative Robot?” marked manually for training the model. Likewise, other questions like “What are the features of CPF, what is CPF, what are the services of Mobile Collaborative Robot” their respective answers are marked manually in the context. These question answers are used as training data for fine-tuning the model.

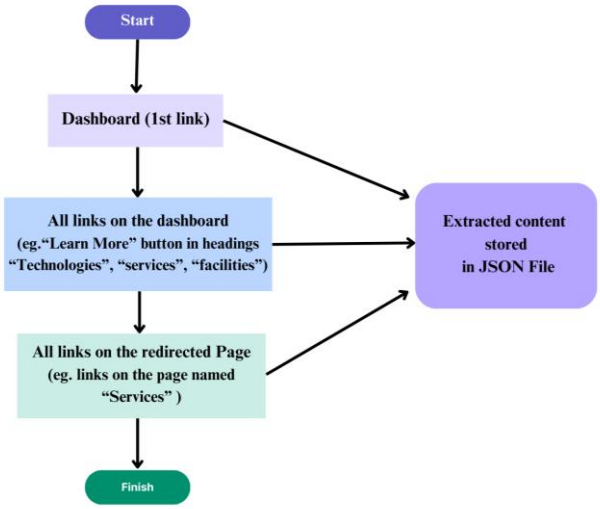

Fig -2: Link Extraction Process

. **Table -1:** Hyper-parameters

Hyper-parameter Tuning	
Learning Rate	3
Max Length	512
Activation	gelu
Stride	100
Batch Size	16
Number of iterations	100
Vocab Size	30522
Model-type	distil-bert

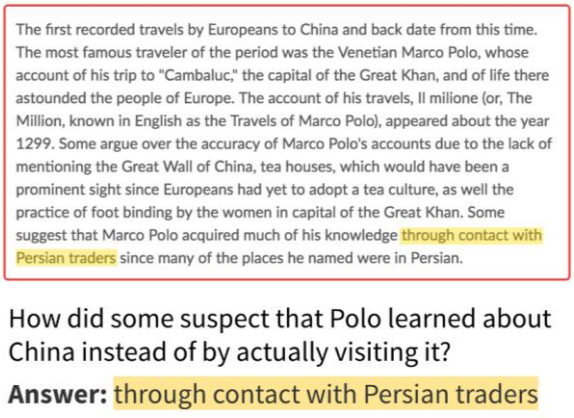

Fig -3: Example of SQuAD

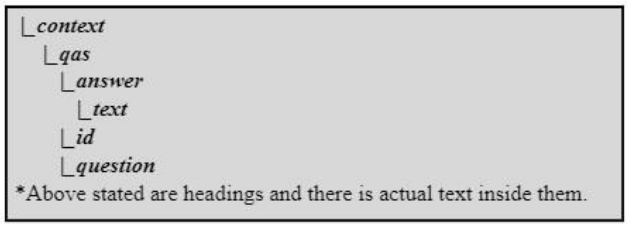

Fig -4: Dataset Structure

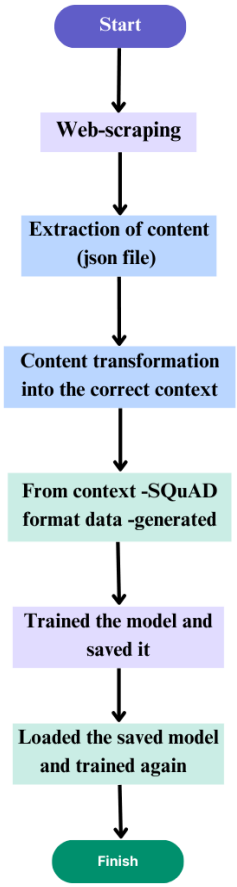

Fig -5: Fine-Tuning Process

4. RESULT

The large language model chat interface prototype has been deployed using the Gradio framework. The model output is compared with the BERT pipeline score and when the score is more than a particular value, the predicted answer is added to the final answer or the predicted answer gives relevance to the question it is added to the final answer. This final answer is the result/chat produced by the model or bot, for example, conversations with the bot in Fig 7 and 8. The BERT model is used for comparison as BERT and Distil-BERT have the same architecture as BERT itself. In order to reduce the size of a BERT model by 40% during the pre-training phase, Distil-BERT knowledge distillation is performed.

3. CONCLUSIONS

This work implements a Chatbot using the NLP to answer queries of users, process parameters, and dataset extraction. Chatbot can answer user queries using transformers, and NLP concepts. Model training is terminated at the loss of 0.048 early-stopping at 312 epochs. It can be concluded that the proposed model benefits the customers; by providing 24x7 support and improving customer service. Chat assistance can solve user queries directly instead of the user exploring the whole website. With the use of chat-bot customer support service will improve a lot where a significant number of questions remain unanswered and hence, unsolved.

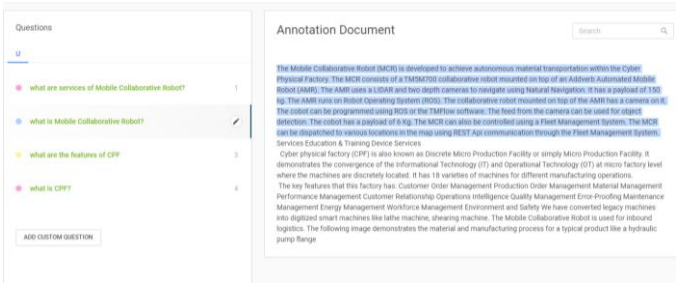

Fig 6: Answering questions about Collaborative Robot

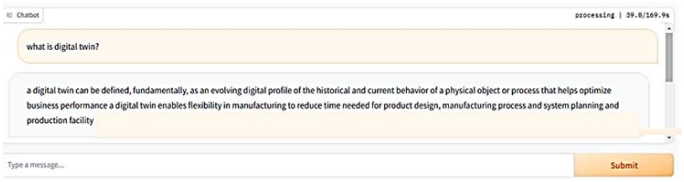

Fig 7: Illustration1

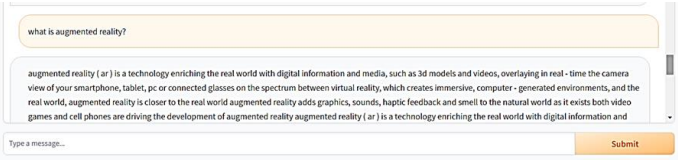

Fig 8: Illustration2

ACKNOWLEDGEMENT

The research presented in this paper was partially supported by the Indian Institute of Delhi - Automation Industry Association - Foundation for Smart Manufacturing (IITD-AIA-FSM) for allowing web-scrapping to collect organizational data.

REFERENCES

[1] M. Akhtar, J. Neidhardt, and H. Werthner, "The potential of chatbots: Analysis of chatbot conversations," in 2019 IEEE 21st Conference on Business Informatics (CBI), vol. 01, pp. 397–404, 2019.

[2] N. Boisgard, State-of-the-Art approaches for German language chat-bot development. PhD thesis, Wien, 2018.

[3] D. Adiwardana, M. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu, and Q. V. Le, "Towards a human-like opendomain chatbot," CoRR, vol. abs/2001.09977, 2020.

[4] M. Nakano and K. Komatani, "A framework for building closed-domain chat dialogue systems," CoRR, vol. abs/1910.13826, 2019.

[5] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," CoRR, vol. abs/1409.3215, 2014.

[6] Y. Liu, M. He, M. Shi, and S. Jeon, "A novel model combining transformer and bi-lstm for news categorization," IEEE Transactions on Computational Social Systems, 2022.

[7] C. Sung, T. Dhamecha, S. Saha, T. Ma, V. Reddy, and R. Arora, "Pre-training BERT on domain resources for short answer grading," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), (Hong Kong, China), pp. 6071–6075, Association for Computational Linguistics, Nov. 2019.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," CoRR, vol. abs/1706.03762, 2017.

[9] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," CoRR, vol. abs/1810.04805, 2018.

BIOGRAPHIES

Nikita is a final year student pursuing B. Tech in Computer Science and Engineering from Jaypee Institute of Information Technology. She has been associated with IITD-AIA-FSM as an ML Intern and has inclination towards research in AI and ML.

Keivalya Pandya is a Mentor in Machine Learning domain at IIT Delhi AIA Foundation for Smart Manufacturing. He is a B. Tech in Mechanical Engineering from Birla Vishvakarma Mahavidyalaya, and serving as a researcher in Robotics, and AI domain.