Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 5, May 2024

AI Adaptable Dynamic Customer Assistance Through Innovative Text Generation

Ms. Swati D. Kadu
Dept. of AI & Ds
Sppu, india

Onkar Sonavane
Dept. of AI & Ds
Sppu, india

Prakalp Pande
Dept. of AI & Ds
Sppu, india

Sushrut Gaikwad
Dept. of AI & Ds
Sppu, india

Yash Jadhav
Dept. of AI & Ds
SPPU, India

***Abstract*--In this paper, we present an innovative approach to customer support systems, focusing on the medical and educational domains, by developing an Adaptable AI Support System for Dynamic Customer Assistance. The purpose of this study was to construct a framework capable of providing dynamic customer assistance across varied sectors and query conditions, with a specific focus on text generation. Methodologically, we utilized natural language processing techniques and transformer-based models to enhance adaptability and effectiveness in customer support systems. Our findings indicate that the integration of attention mechanisms and fine-tuning strategies significantly improves the precision and F1 score in textual data generation. The practical implications of this research lie in the development of more responsive and contextually relevant customer support systems tailored to specific domains. The originality of this work lies in its comprehensive methodology and the exploration of advanced techniques in text generation.**

I. INTRODUCTION

In recent years, the landscape of customer support systems has been rapidly evolving, driven by advancements in artificial l intelligence (AI) and natural language processing (NLP) technologies. Traditional customer support methods, reliant on human operators, have increasingly given way to automated systems capable of interpreting and responding to user queries in real-time. However, while these systems have demonstrated efficacy in certain contexts, they often lack the adaptability required to handle the dynamic and diverse nature of customer interactions across different sectors and query conditions. This paper presents a novel approach to address this challenge through the development of an Adaptable AI Support System for Dynamic Customer Assistance. The primary objective of this research is to construct a framework capable of providing responsive and contextually relevant assistance across varied sectors, with a particular emphasis on the medical and educational domains.

To achieve this objective, we adopt a comprehensive methodology that integrates cutting-edge NLP techniques and transformer-based models. Unlike previous approaches that focus on static responses or domain-specific solutions, our framework prioritizes adaptability and effectiveness in handlinga wide range of user queries. The methodology and implementation section of this paper detail our approach, starting with the collection and preprocessing of diverse datasets encompassing a broad spectrum of language patterns and topics within the targeted domains. We then employ advanced techniques such as word embedding, recurrent neural networks (RNNs), and attention mechanisms to enhance the adaptability and effectiveness of our model in generating contextually relevant responses.

A key contribution of this research lies in the exploration of attention mechanisms and fine-tuning strategies, which significantly improve the precision and F1 score in textual data generation. By dynamically assigning weights to words based on their importance, our model can generate more relevant and contextually coherent responses to user queries. The findings from our experiments demonstrate the efficacy of our approach, with our model outperforming existing baselines across various metrics. The practical implications of this research are significant, as it lays the groundwork for the development of more responsive and adaptable customer support systems tailored to specific domains. This paper contributes to the ongoing discourse on AI-driven customer support systems by introducing a novel framework that prioritizes adaptability and

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 5, May 2024

effectiveness. Through a combination of advanced NLP techniques and transformer-based models, we demonstrate the potential for enhancing customer assistance across diverse sectors and query conditions. The subsequent sections of this paper provide a detailed exploration of our methodology, findings, and implications, offering valuable insights for researchers and practitioners in the field.

II. METHODOLOGY AND IMPLEMENTATION

In this section, we configure the methodology to construct an Adaptable AI Support Systems for Dynamic Customer Assistance Across Varied Sectors and Query Conditions focusing on text generation. Our main objective is to build a framework which can provide dynamic customer assistance under various domains, focusing on medical and educational sectors.

As the existing works have effective result in text generation,our approach is to enhance the adaptability and effectiveness of systems. With the help of the natural Language Processing techniques and transformer-based models, we seek to create an innovative solution for customer support that goes beyond conventional methods. The workflow depicted in Fig. 1, shows a visual representation of our model approach in building the adaptable AI customer support system.

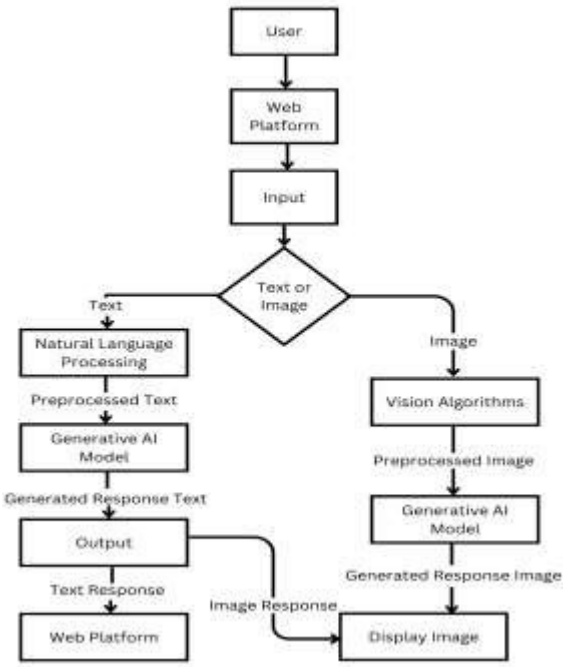

Fig .1 Workflow for the Adaptable AI Support System

With the help of this comprehensive methodology, we strive to provide the advanced technology of text generation support systems. The other sections such as utilization of transformer- based models, attention mechanisms and fine-tuning and hyperparameter optimization.

A. Problem Assertion

The model addressed the problem of building an adaptable AI support system based on different domains, mainly on medical and educational domains. The project addresses the need of the text-based interaction of the project dynamically generating the textual information relevant to the user's query by specialized fields. The model aims to grasp advanced technologies like transformer-based models to overcome the limitations of the conventional methods, creating a diversified and adaptable AI support system that generates the textual data precision in response generation.

B. Data Collection

To enhance the model training, we employ the substantial dataset encompassing a diverse customer query and their responses. Each query, denoted as $\boldsymbol{Q}$, and its corresponding response, denoted as $\boldsymbol{R}$, constitute individual training instances. The dataset encapsulates a broad spectrum of language patternsand topics within the targeted domains.

C. Data Preprocessing

In the data preprocessing step [12], we standardize the input texts by transforming it into lowercase and performing the tokenization process followed by removing special characters. The process eliminates the extraneous components, resulting inrefined data. Further stemming and lemmatization process of reducing inflected words to their word stem. After this, Vectorization plays its role with Bag of Words (BOW) and Term Frequency and Inverse Document Frequency (TF-IDF) by converting terms into mathematical or vector format which is understandable by machines. In Bag of Words (BOW), the presence of words is written by 1 and absence by 0. As there is a semantic meaning of Bag of Words, so TF-IDF overcomes it by providing a crucial aspect as, higher the TF-IDF value, morethe important semantic meaning of the word.

D. Word Embedding

In Natural Language Processing (NLP), the general term for language models and representation learning techniques is Word embedding [1][13] for text analysis, typically in the form of a real-valued vector. In the project, word embedding technique of Skip-Gram is applied. This enhances the training of the model based on the textual data and increases the ability to understand and generate the relevant responses of the queries. The

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 5, May 2024

word2vec algorithm uses a neural network model to learn the word associations from a large amount of the textual data. Oncethe model is trained, it can detect synonymous words or can suggest additional words for a partial sentence. As the name implies, word2vec represents each distinct word with a particular list of numbers said to be a vector. The skip-gram model captures the semantic and syntactic relationships between the words to make them prepared for NLP related tasks. Other techniques like GloVe, ConceptNet Numberbatch are also used for the word embedding task. So, the utilization of this technique word embedding played an important role in our project, allowing the model to precisely capture the semantic relationships between the textual data and their contextual usage.

E. Recurrent Neural Networks

a. Sequence to Sequence Models

The Sequence-to-Sequence (Seq2Seq) model operates by taking an input sequence, processing it, and generating an output sequence, making it particularly effective in text generation. The Seq2Seq architecture consists of two components: an encoder and a decoder.

Our project leverages a Seq2Seq [3] model to build a conversational AI. This model takes user input and generates responses. We have trained the model on massive dialogue datasets; the model aims to provide natural responses across different topics.

During the training phase, the encoder processes the input sequence and generates a context vector, then decoder takes this context vector and generates the output sequence. This process involves the use of transformer architectures for both the encoder and decoder, allowing the model to generate output sequences.

Let input sequence $(x_1, x_2, x_3, \ldots, x_T)$ and transformer generates output sequence $(y_1, y_2, y_3, \ldots, y_T)$

$$h_t = \sigma(\omega^{hx} x_t + \omega^{hh} h_{t-1})$$

where,

h_t hidden state at the time t,

x_t represents the input at time t,

ω^{hx} represents the weight matrix for the input,

ω^{hh} represents the weight for the hidden state from previoustime (t-1),

σ represents the activation function.

$$y_t = \omega^{yh} h_t$$

where, y_t represent the output at time, ω^{yh}represents the weight matrix for the output,

h_t represents the hidden state at time t from the encoder ordecoder.

b. Bidirectional RNN

Bidirectional RNNs (Recurrent Neural Networks) [7] possess the power to handle the techniques for text generation. As compared to traditional RNNs, those processes textsequentially, but Bidirectional RNNs analyse the text considering both the future and the past. It helps the current context to capture more correct relevance from future words andprovides richer context and it enables more grammatically enhanced text generation.

In our project, Bidirectional RNNs [11] are applied for text generation tasks as a relevant response to a query. It allows the model to catch information from subsequent and preceding tokens in the provided input sequence. So, this leverages the comprehensive understanding of the context, leading to enhanced text generation performance.

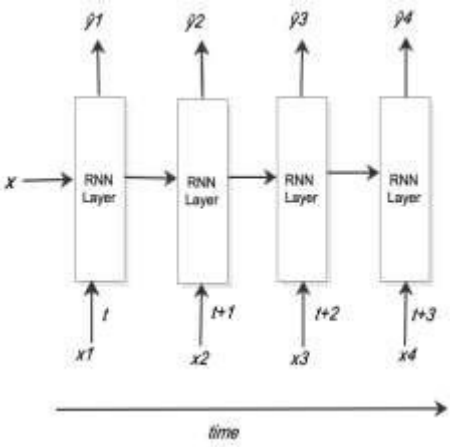

Fig.2 Unidirectional RNN

In some cases, the outcome of one hidden layer is dependent onfuture outcomes. As shown in figure 1, in case of Unidirectional RNN, for each input words *x* passed to the RNNlayer, output of *ŷ3* can be dependent on *ŷ4,* so output of layer 3may not be able to generate expected result because of the lack of information which is upcoming information. This problemcan be solved by Bidirectional RNNs.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 5, May 2024

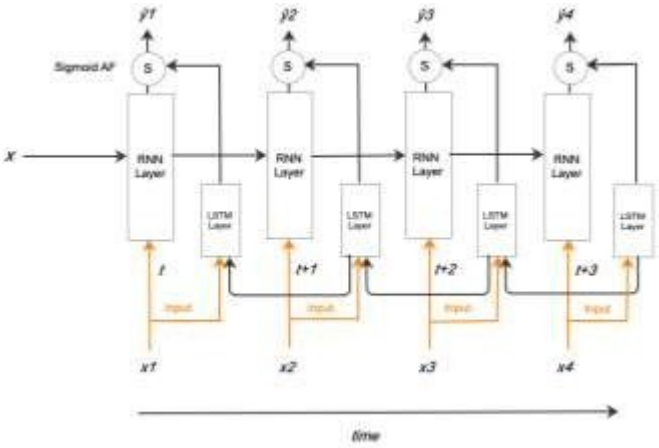

Fig.3 Bidirectional RNNs

c. LSTM RNNs

Long Short-Term Memory (LSTM) [2] these networks are the special type of RNNs. This architecture is developed for long-term dependencies in the textual data or sequential data. They perform well for tasks like text generation as they keep the whole information which is important through the whole sentence. LSTM RNNs have the ability to retain and utilize the textual data over an extended period of time generating more corresponding and relevant and contextually relevant data.

In our project, LSTM RNNs [9] played an important role ingenerating textual responses to the user's queries. LSTM RNNs were trained on huge text corpora to adapt the hidden patterns and architecture of the language, giving them privilege to generate meaningful and corresponding relevant text outputs.

LSTM RNN networks consist of memory cells that maintain a call state and other three types of gates: input gate, forget gate, output gate. With help of these gates, the flow of the information in and out of the memory cells, as this activates the LSTM networks to selectively retain or lose the information over the period. This architecture allows the LSTM RNNs to catch the long-term dependencies in sequential data. Resulting in more effectiveness for textual data generation.

d. RNN Encoder-Decoder

The RNN Encoder-Decoder [4] structure is used in natural language processing tasks which is framework, also contributesto text generation. It Consists of two recurrent neural networks (RNNs) - one of them is encoder and other one is decoder. These two encoder and decoder perform together to generate the text.

The encoder works on the input sequences and vectorized fixed-dimensional conversion is represented. Based on the vectorized representation, the decoder yields the output sequences.

In our project, we applied RNN Encoder-Decoder [5] architecture for generating the text in context of customer support system, mainly focusing on medical and education domains. The queries or input prompts are processed by the encoder network and the decoder network creates the corresponding related output responses. To achieve the accuracy of the responses, we trained the model on huge dataset of large corpus of related and specific data to each topic so that we ensured that we get contextually coherent and relevant responses.

Arithmetic representation for Encoder and Decoder

$$h_t = f_{enc}(x_t, h_{t-1}) \quad (1)$$

$$c = q_{enc}(h_1, h_2, \ldots, h_T) \quad (2)$$

$$h' = f_{dec}(y_{t-1}, s_t) \quad (3)$$

$$p(y_t, y_{t-1}, \ldots, y_1, c) = g(y_{y-1}, h', c) \quad (4)$$

$$Loss = -\frac{1}{N}\sum_{i=1}^{N} \log p(y_i \ldots, c) \quad (5)$$

Equation (1) represents the recurrent step in encoder network, computing hidden states.

Equation (2) represents the aggression of the encoder hidden states into a context vector.

Equation (3) represents the step in decoder network, generatinghidden states.

Equation (4) represents the probability distribution over output tokens given context.

Equation (5) represents the loss function measuring model's prediction accuracy.

e. Attention Mechanism:

Attention mechanisms have majorly contributed to Natural Language Processing (NLP) by activating the models to give more attention to a specific part or those parts which has more importance in the process of generating the output sequenced response. In the attention mechanism, each word is assigned with some weights dynamically, which represents the importance of that word while producing each word in the response. By assigning weights, richer and more relevant and contextually coherence textual data which emphasise the important words.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 5, May 2024

In or project, to get more relevant responses, we have integrated attention mechanism into the transformer-based architecture, so it helps to give more relevant responses to the queries. In addition, we have fine-tuned pre-trained transformer models which is trained on specific data domains to improve accuracy and the responses. The multi-head attention mechanism, it performs attention computations simultaneously, so helps model to catch the various aspects of the input query.

$$H = w_1 x_i + b_1 \quad (6)$$

$$M = \tanh(H) \quad (7)$$

$$\alpha = soft\max(w^T M) \quad (8)$$

$$R = H\alpha^T \quad (9)$$

$$h = \tanh(R) \quad (10)$$

Equation (6) represents the new weight distribution layer.
Equation (7) represents the linear transformation with tanh activation function
Attention layer I includes the equations from (6) to (10)
Attention layer II includes the equations from (7) to (10)

f. Fine Tuning

Fine-tuning [6][10] in context of the text generation this technique refers of adapting a pre-trained language model to a specifies domain by training it with the domain-specific data. This enhances the model to learn the nuances and complexities of the intent domain, leading to improving the performance also generating the relevant text.

In our project, we applied the fine-tuning methods to adjust and customize pre-trained transformer based large languagemodels for customer support systems in domains such as medical and education, etc. For the specific domains, gathered the dataset of the required domains and fine-tuned pre-trained models like BERT [8], GPT-3, Seq2Seq on textual data which leads to make them proficient in producing contextually relevant responses for customers input queries.

III. RESULTS AND FUTURE WORK

In the implemented experiment, different recurrent neural network (RNNs) architectures were measured with their performance in the response generation. The models we evaluated includes LSTM, GRU, Transformer, GPT, Seq2Seq and BERT. The LSTM and GRU models use the traditional choices for sequential text generation, using gates input, memory and predictions. LSTM uses the three gates (input, forget and output). GRU model employs reset and update. Transformer models like GPT, implies multi-head attention mechanisms. BERT, a pre-trained transformer model, enhances the context by bidirectional training. Seq2Seq models, most basic architecture gives the bottom-line comparison.

TABLE.1 COMPARISON BETWEEN VARIOUS ARCHITECTURES

Model	Epoch	Learning Rate	Precision	F1
LSTM	5	0.001	0.85	0.82
GRU	5	0.001	0.87	0.84
Transformer	5	0.001	0.90	0.88
GPT	5	0.001	0.88	0.86
Seq2Seq	5	0.001	0.86	0.83
BERT-based Encoder-Decoder	5	0.001	0.93	0.91

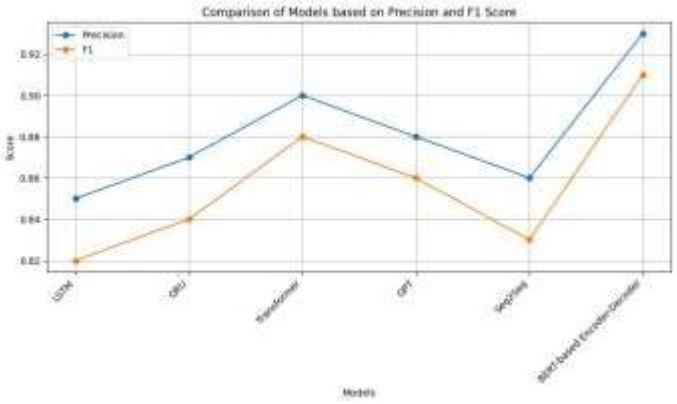

Fig.4 Comparison between various models

IV. CONCLUSION

In this paper, we propose an innovative approach to customersupport systems, focusing on domains like medical and education, with development of an Adaptive AI Support System for Dynamic Customer Assistance. Our implementations with different recurrent neural network architectures, including LSTM, GRU, Transformer, GPT, Seq2Seq and BERT models, we compare the effectiveness of these models demonstrating the high precision and F1 score in textual data generation across specifies domains. The exploration of attention mechanisms and fine-tuning strategies shows their important role in leveraging the relevant responses. From the experimental results, we can assume the attention model architectures have done remarkable effects in text generation. In future work, we seek to apply this to solve more complex problems providing high precision. Therefore, we conclude that the BERT-based Encoder-Decoder model demonstrates the highest efficacy in our study and holds

promise for future applications requiring high precision in text generation tasks.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 5, May 2024

ACKNOWLEDGMENTS

We would like to sincerely thank Ms. Swati D. Kadu for her invaluable guidance and support as our guide. We also extend our appreciation to our colleagues for their collaboration and assistance throughout this research endeavour.

REFERENCES

[1] D. Purwitasari, A. A. Zaqiyah and C. Fatichah, "Word-Embedding Model for Evaluating Text Generation of Imbalanced Spam Reviews," 2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Depok, Indonesia, 2021, pp. 1-6, doi: 10.1109/ICACSIS53237.2021.9631315.

[2] H. V. K. S. Buddana, S. S. Kaushik, P. Manogna and S. K. P.S., "Word Level LSTM and Recurrent Neural Network for Automatic Text Generation," 2021 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2021, pp. 1-4, doi: 10.1109/ICCCI50826.2021.9402488.

[3] S. Zhao, E. Deng, M. Liao, W. Liu and W. Mao, "Generating summary using sequence to sequence model," 2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC), Chongqing, China, 2020, pp. 1102-1106, doi: 10.1109/ITOEC49072.2020.9141919.

[4] M. Kumar, A. Singh, A. Kumar and A. Kumar, "Analysis of Automated text generation using Deep learning," 2021 Fourth International Conference on Computational Intelligence and Communication Technologies (CCICT), Sonepat, India, 2021, pp. 14-18, doi: 10.1109/CCICT53244.2021.00014.

[5] C. Çağlayan and M. Karakaya, "Topic-Controlled Text Generation," 2021 6th International Conference on Computer Science and Engineering (UBMK), Ankara, Turkey, 2021, pp. 533-536, doi: 10.1109/UBMK52708.2021.9558910.

[6] M. Rehan, M. S. I. Malik and M. M. Jamjoom, "Fine-Tuning Transformer Models Using Transfer Learning for Multilingual Threatening Text

Identification," in IEEE Access, vol. 11, pp. 106503-106515, 2023, doi: 10.1109/ACCESS.2023.3320062.

[7] B. Wang, F. Miao, X. Wang and L. Jin, "Text Classification Using a Bidirectional Recurrent Neural Network with an Attention Mechanism," 2020 International Conference on Culture-oriented Science & Technology (ICCST), Beijing, China, 2020, pp. 265-268, doi: 10.1109/ICCST50977.2020.00057.

[8] S. Wu, Z. Huang and H. Feng, "Text Labels Classification Model based on BERT Algorithm," 2023 4th International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE), Nanjing, China, 2023, pp. 329-332, doi: 10.1109/ICBASE59196.2023.10303262.

[9] S. Chakraborty, J. Banik, S. Addhya and D. Chatterjee, "Study of Dependency on number of LSTM units for Character based Text Generation models," 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA), Gunupur, India, 2020, pp. 1-5, doi: 10.1109/ICCSEA49143.2020.9132839.

[10] T. H. Daryanto and M. L. Khodra, "Indonesian AMR-to-Text Generation by Language Model Fine-tuning," 2022 9th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA), Tokoname, Japan, 2022, pp. 1-6, doi: 10.1109/ICAICTA56449.2022.9932960.

[11] H. Asrawi, A. Sunyoto and B. Setiaji, "Implementation of Bidirectional Gated Recurrent Units for Text Classification," 2023 6th International Conference on Information and Communications Technology (ICOIACT), Yogyakarta, Indonesia, 2023, pp. 464-469, doi: 10.1109/ICOIACT59844.2023.10455822.

[12] Yuyun, A. D. Latief, T. Sampurno, Hazriani, A. O. Arisha and Mushaf, "Next Sentence Prediction: The Impact of Preprocessing Techniques in Deep Learning," 2023 International Conference on Computer, Control, Informatics and its Applications (IC3INA), Bandung, Indonesia, 2023, pp. 274-278, doi: 10.1109/IC3INA60834.2023.10285805.

[13] A. Neelima and S. Mehrotra, "A Comprehensive Review on Word

Embedding Techniques," 2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS), Coimbatore, India, 2023, pp. 538-543, doi: 10.1109/ICISCoIS56541.2023.10100347.