https://africanjournalofbiomedicalresearch.com/index.php/AJBR
Afr. J. Biomed. Res. Vol. 27 (October 2024); 908-912

Research Article

TherapEase: Conversational Chat-bot for Mental Health Screening Using Trained Transformer

Dr. Manish Rana[1*], Dr. Mahendra S. Makesar[2], Prof. D.R. Solanke[3], Mr. Suresh R. Mestry[4], Dr. Sunny Sall[5], Prof. Devki Nadgaye[6]

[1]Associate Professor of Information System, St. John College of Engineering & Management (SJCEM) Palghar-401404, INDIA. E-Mail: dr.manish_rana@yahoo.co.in.*
[2]Associate Professor of Information Technology, Nagpur Institute of Technology (NIT), Nagpur, Maharashtra - 441501, INDIA. E-Mail: mahendramakeshwar@gmail.com.
[3]Assistant Professor of Applied Electronics, Sant Gadge baba Amravati University, Amravati(SGBAU)-Amravati-444602, INDIA.E-Mail: solankedr@gmail.com.
[4]Assistant Professor of Computer Engineering, Rajiv Gandhi Institute of Technology (RGIT), Andheri Mumbai-400061, INDIA. E-Mail: suresh.mestry@mctrgit.ac.in.
[5]Assistant Professor of Computer Engineering, St. John College of Engineering & Management (SJCEM) Palghar-401404, INDIA. E-Mail: sunny_sall@yahoo.co.in.
[6]Assistant Professor of Information Technology, Nagpur Institute of Technology (NIT), Nagpur, Maharashtra - 441501, INDIA. E-Mail: devki .nandhaye@gmail.com.

Abstract:
The present research explores the pre-trained transformer's capability to provide conversational chatbot, using a niche approach of fine tuning the transformer on a custom dataset. Rags, Sentence Transformers, LLMs are all part ofthis suggested technique's implementation. This experiment examines the versatility of a pre-trainedtransformer and its ability to not only provide semantic context to highly sensitive conversationtopics but also use its pre-training knowledge base to provide appropriate recommendations. Theresults are more promising to increase the nuances of a conversation chatbot. We give an original perspective on increasing the accuracy and effectiveness of an advanced chatbot that exceeds existing methodologies, revealing insight on the expanding landscape of artificial conversational chatbots using this complete methodology.

Keywords: Hybrid Technique, Fine Tuning, Custom Dataset, Transformers, Large Language Models, Retrieval Augmented Generation (RAG), Mental Health Chatbot, BERT, Hugging Face

Receiving Date: 27/09/2024 Acceptance Date: 16/10/2024

DOI: https://doi.org/10.53555/AJBR.v27i3.2082

INTRODUCTION

The rise in technological advancements has led to the invention of automated tasks that ease up human labour and are more efficient in the long run. Chatbot is one ofthose inventions which cater to needs in areas where a human presence is not feasible. A chatbot is a computer program that simulates and processes human conversation which thereby allows humans to interact with digital devices as if they were communicating witha real person.

Chatbot serves multiple use cases, anything from answering frequently asked question to serving a full- length conversation. The technology that is used to design and deploy a chatbot has changed drastically over the years. Earlier, the chatbots were made with a focus on rule-based systems and pre-defined responseswhere developers defined a set of rules and responses for the same. These rules were often based on keywordmatching

which would give a dedicated answer if the user entered a keyword pre-determined in the knowledge base.

Due to the limited advancements in NLP at the time, there were only a certain thing that a chatbot was designed to do like answering FAQs and reading out some important details regarding a product or a servicefor which hiring a human employ wouldn't be economically feasible.

Limitations of these early chatbots were limited understanding of context, dependency on keywords andthe lack of learning abilities.
Due to the rapid advancements in machine learning, weare at a stage that the limitations of the previous chatbots are overcome with the help of algorithms like LSTM and Transformers.

These algorithms are valuable as they help us in storing the semantic relation between sentences for a longer duration of time thereby increasing the understanding of context and ability to answer the question asked more accurately. These advanced neural networks models excel at capturing sequential information, handling long-term dependencies, and incorporating sophisticated attention mechanisms. These capabilities are crucial in chatbot applications were understanding context, maintaining conversational flow, and generating coherent and contextually relevant responses are essential for enhancing accuracy and overall performance

LITERATURE REVIEW

The limits of conventional sequence-to-sequence models, which analyse sequential data using recurrence and convolution, are first highlighted in the work. The authors contend that these models are not ideal for tasks like text summarization and machine translation that call for long-range interdependence and contextual comprehension.
Context: The writers give a succinct synopsis of sequence-to-sequence assignments and the several methods that have been suggested for solving them. They talk about the drawbacks of conventionaltechniques, such as the need for convolution and repetition, as well as the challenge of training these models.
Self-Attention Mechanism: As an alternate method of processing sequential input, the authors present the self-attention mechanism [1].

A new paradigm called Federated Learning (FL) allows users who are geographically dispersed to train machine learning models together and iteratively without exchanging private information. Driven by the efficiency and resilience of self-attention-based structures, scientists are resorting to use pre-trained Transformers, also known as foundation models, in FL rather than conventional convolutional neural networks, in order to take advantage of their superior transfer learning qualities. Despite recent advancements, it is still unclear how pre-trained Transformer models fit into FL—that is, how FL userscan profit from this new paradigm and how to effectively refine these pre-trained models in FL. In this study, we investigate this problem and show that the Transformers that have been fine-tuned perform exceptionally well on FL, and that the lightweight fine-tuning approach enables a quick convergence rate andminimal communication overhead [2].

TRANSFORMERS

This paper provides an introductory exploration of transformers, a revolutionary neural network architecture, and their applications in chatbot technology. Transformers, introduced in the seminar paper "Attention is All You Need," have transformed the landscape of natural language processing (NLP) andconversational AI.
Transformers consist of an encoder-decoder architecture, with multiple layers of self-attention mechanisms and feedforward neural networks. Theencoder processes the input text, while the decoder generates the output text, each layer in the transformer architecture processes the input sequentially, allowing the model to learn complex patterns and relationships within the data [1].

To build our chatbot we used BERT [Bidirectional Encoder Representations from Transformers], which a special type of transformer architecture design is specially catering to the natural language processing (NLP) related applications.
The reasons why we opted for using BERT are as follows:

1) Due to its bi-directional nature, it can predict the semantic context and voice of the sentence very accurately.
2) It provided contextual word embedding's which can be helpful to understand the seriousness of a user and give an appropriate suggestion.
3) High accuracy because it is trained on a very largedataset that includes multiple languages. It excels in general language pattern, structures, andrelationships of the sentences.

How does BERT actually know what solution tosuggest?
RAG :(Retrieval-Augmented Generation)is an approach that combines a retriever and a generator in aunified model. The retriever selects relevant passagesfrom a large dataset, and the generator utilizes thesepassages to generate the final response. It helps BERTencode passages from a large dataset on which it is pre-trained. The embedding's that are generated arecompared based on the relevance of a given sentence. RAG is a seq2seq model which encapsulates two core components: a question encoder and a generator.During a forward pass, we encode the input with thequestion encoder and pass it to the retriever to extractrelevant context documents. The documents are then prepended to the input. Such contextualized inputs arepassed to the generator. The model can be initialized with a Rag-Retriever for end-to-end generation or usedin combination with the outputs of a retriever inmultiple steps---see examples for more details. The model is compatible any *auto encoding* model as the question encoder and any *seq2seq* model with languagemodel head as the generator. It has been tested with DPR Question Encoder as the question encoder and Bart For Conditional Generation or T5 For Conditional Generation as the generator [3].

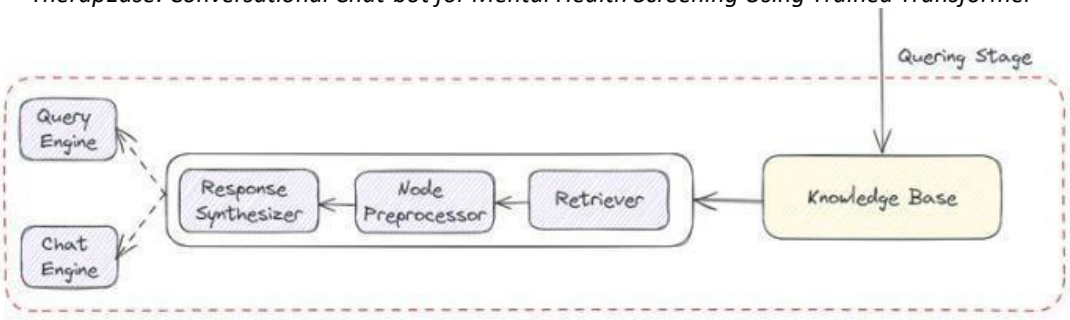

Figure 1 -BERT IMPLEMENTATION

METHODOLOGY

A pre-trained transformer is a specific kind of transformer neural network designed for tasks like natural language processing. During pertaining, it learns from a large dataset, gaining a generalunderstanding of language patterns. Afterward, it's fine-tuned for a specific task, adapting its knowledge for more efficient performance.

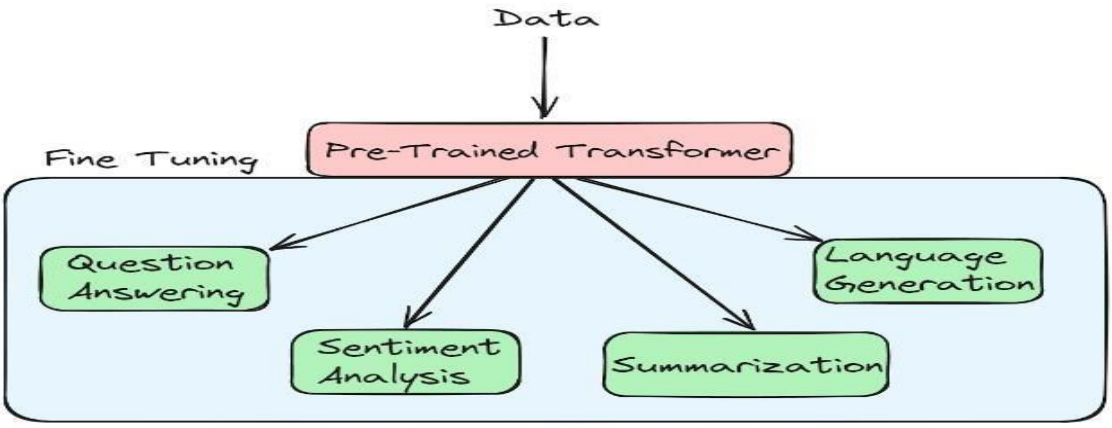

Figure 2 - Role of Pre-Training and Fine Tuning

These pre-trained transformers are a subtype of transformers, utilising transfer learning. Transfer learning involves training a model on one task and applying that knowledge to a related task. For instance, think of training a convolutional neural network (CNN)on ImageNet. The CNN can then be fine-tuned to classify specific types of vehicles, like distinguishing between a Tesla and an SUV. The initial training on ImageNet provides a foundation for recognizing general image features, which is then refined for the new task [4].

However, We wanted to take this one step further and we used BERT to fine tune it on a custom dataset. Finetuning is a process in machine learning where a pre- trained mode (BERT), further training it on a smaller, task specific dataset to adapt its knowledge to a specific application. Fine-tuning allows the model to leverage the general knowledge it gained in its pre-training and tailor it to a more nuanced mental health analysis setting. The method to fine tune BERT on a custom mental health dataset was to check the sentiment of thesentence and give suggestions for the same [5].

IMPLEMENTATION

In implementing our mental health chatbot, we harness the power of BERT, a state-of-the-art transformer model known for its robust contextual understanding abilities. BERT will serve as the backbone for natural language understanding, enabling the chatbot to grasp the intricate nuances of user queries, detect emotions, and provide empathetic responses.

To streamline the user interface and dialogue management, we leverage Stream lit, a user friendly Python library. Stream lit allows us to create an interactive and visually appealing environment, facilitating seamless communication between users andthe chatbot. It also enables the storage of conversation history, ensuring continuity and context-aware responses throughout the interaction.

The heart of our mental health chatbot lies in BERT's bidirectional contextual understanding. This capability allows the model to maintain a comprehensive view ofthe conversation, ensuring that user inputs are interpreted within the appropriate context. As users share their thoughts and feelings, BERT's contextual awareness aids in generating responses that are not onlyaccurate but also sensitive to the evolving emotional state of the user [6].

For user engagement and model refinement, we implement a feedback system. Users have the opportunity to provide feedback on the chatbot's responses, indicating the effectiveness and relevance of the provided support. This valuable feedback will be systematically collected and analyzed. Periodic surveys and sentiment analysis on user interactions willcontribute to refining the model, making it more adeptat offering accurate suggestions and mental health tipsover time.

Deploying this sophisticated mental health chatbot onthe web is made seamless through Hugging Face's Transformers library. Hugging Face provides a robustplatform for model deployment, ensuring accessibilityto a wide audience. By utilizing Hugging Face's deployment capabilities, we make our chatbot readily available to users seeking mental health support online, breaking barriers and helping where it's neededmost.

In summary, our mental health chatbot integrates the strengths of BERT for contextual understanding,Stream lit for interactive dialogue management, and Hugging Face for efficient model deployment. The fusion of these technologies creates a supportive and user friendly environment that not only addresses mental health concerns effectively but also continuously evolves through user feedback, ensuring a compassionate and tailored experience for each interaction [7][8].

CONCLUSION

In conclusion, the implementation of our mental health chatbot represents a powerful fusion of cutting-edge technologies. By harnessing the contextual understanding abilities of the BERT pre-trained model, fine-tuned on a custom dataset tailored to mental healthconcerns, we ensure a nuanced and empathetic engagement with users. BERT's bidirectional context comprehension enables our chatbot to provide personalized and sensitive responses, fostering asupportive environment for users seeking mental healthassistance.

By seamlessly integrating these technologies, our mental health chatbot endeavors to make a meaningful impact, offering support, understanding, and valuable resources to those navigating the complexities of mental wellbeing. As technology advances, so does our commitment to leveraging it for the betterment of mental health, fostering a future where compassionate and tailored support is easily accessible to all.

LIMITATIONS

The inherent nature of this BERT trained model presents challenges in delivering concise answers to user queries. Due to its design, there is a necessity for some level of hard coding to ensure clear and effective responses, thereby enhancing user engagement. However, this approach might not cover all possible scenarios, leading to potential gaps in understanding niche contexts that lack sufficient representation in thetraining data. Consequently, the model may generate suggestions that are irrelevant or impractical.

To address this limitation, there is a crucial need to prioritize locality in the model's training. Yet, this introduces a trade-off with privacy and security concerns, as increased focus on local data may expose sensitive information. Striking the right balance between localized knowledge and safeguarding user privacy becomes imperative.

Furthermore, users are required to provide detailed explanations of their problems and surrounding context for the model to generate meaningful suggestions. This could be cumbersome, especially in an era where attention spans are diminishing, and users seek quick and efficient solutions.

Continuous improvement is essential, necessitating a constant loop of explicit and implicit feedback. Users' input becomes instrumental in refining the model'sperformance over time. This iterative feedback processis crucial for addressing its limitations and enhancing its ability to adapt to evolving user needs.

FUTURE SCOPE

The forthcoming pivotal focus lies in designing a well-crafted user interface, leveraging either Flask or Next JS. Transitioning this prototype model into a production environment not only holds the promise of attracting a larger user base but also serves as a crucialavenue for assessing the model's robustness. The deployment strategy leans towards AWS, given its scalability and reliability, although it's worth noting that AWS usage incurs costs per inference. To alleviate financial constraints, a concerted effort will be directed towards fundraising and outreach initiatives.

For a comprehensive evaluation and ongoing testing, the implementation of generative prompt transformer models like LLM through Lang chain emerges as avaluable approach. These models can generate visual graphs, enhancing the credibility of the suggestions provided. Adding another layer of functionality, a GPTmodel could be integrated to generate concise summaries or receipts summarizing each full-lengthconversation a user has with the system. This multi- modal approach enriches the user experience and provides tangible outputs for reference.

In essence, the envisioned production setting involves a synergistic integration of a refined user interface, AWS deployment, financial support mechanisms, and advanced generative models to elevate the model's performance and user interaction to new heights.

ACKNOWLEDGEMENT

This research was supported/partially supported by [Dr. Manish Rana, Dr. Sunny]. We thank our colleagues from [St. John College of Engineering and Management who provided insight and expertise that greatly assisted the research, although they may not agree with all of the interpretations/conclusions of this paper. We thank [Dr. Mahendra S. Makesar, Prof. Devki Nadgaye from (Nagpur Institute of Technology (NIT), Nagpur] & [Prof. Darasing Ramrao Solanke from Sant Gadge baba Amravati University, Amravati (SGBAU)]- for assistance with [Cognicraft, Decision-making Techniques], and [Mr. Suresh R. Mestry, Assistant professor, Rajiv Gandhi Institute of Technology (RGIT), Andheri] for theoretical Significance that greatly improved the manuscript.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[2] J. Chen et al., “FedTune: A deep dive into efficient federated fine-tuning with pre-trained transformers,” *arXiv preprint arXiv:2211.08025*, 2022.

[3] P. Lewis et al., “Retrieval-augmented generation for knowledge-intensive NLP tasks,” *Advances in Neural

Information Processing Systems*, vol. 33, pp. 9459-9474, 2020.

[4] R. Doi, T. Charoenporn, and V. Sornlertlamvanich, "Automatic question generation for chatbot development," in *2022 7th International Conference on Business and Industrial Research (ICBIR)*, Bangkok, Thailand, 2022, pp. 301-305.

[5] Z. Liu et al., "Improved fine-tuning by better leveraging pre-training data," *Advances in Neural Information Processing Systems*, vol. 35, pp. 32568-32581, 2022.

[6] J. Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[7] K. N. and U. J., "MediBot: Healthcare assistant on mental health and well-being," in *2023 7th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, Bangalore, India, 2023, pp. 15.

[8] N. Anand, L. M. Pant, T. Alam, S. Pundir, L. Thomas, and U. R. Rakshith, "Artificial intelligence's contribution to mental health education," in *2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, Erode, India, 2023, pp. 462-467.

[9] R. Crasto, L. Dias, D. Miranda, and D. Kayande, "CareBot: A mental health chatbot," in *2021 2nd International Conference for Emerging Technology (INCET)*, Belagavi, India, 2021, pp. 1-5.

[10] L. Brocki, G. C. Dyer, A. Gładka, and N. C. Chung, "Deep learning mental health dialogue system," in *2023 IEEE International Conference on Big Data and Smart Computing (BigComp)*, Jeju, Republic of Korea, 2023, pp. 395-398.