Sahara: Virtual Companion - Exploring Multimodal Empathetic Conversational AI leveraging Ensemble Learning with Humor

Smit D. Gaikwad[1], Aishwarya B. Iyer[2], Krishna S. Talluri[3], Pradip Salve[4]

[1]School of Engineering, Ajeenkya D.Y Patil University, Charholi Budruk, Pune
Email: *smit.Gaikwad[at]adypu.edu.in*

[2]School of Engineering, Ajeenkya D.Y Patil University, Charholi Budruk, Pune
Email: *Aishwarya.Bhaskar[at]adypu.edu.in*

[3]School of Engineering, Ajeenkya D.Y Patil University, Charholi Budruk, Pune
Email: *Talluri.Sai[at]adypu.edu.in*

[4]School of Engineering, Ajeenkya D.Y Patil University, Charholi Budruk, Pune
Email: *pradip.Salve[at]adypu.edu.in*

Abstract: ***Exchanging conversations for increasing mindfulness and combatting loneliness helps reduce and prevent mental health problems. Anxiety, depression, and stress among other mental health disorders are growing in today's world that is driven by speed and changes, severely affecting the lives and well-being of people across all ages and diversities. The present research aims to design a virtual companion which resolves problems with an inclusive approach, is aimed at looking at a problem through different perspectives to get a holistic understanding, using the best natural language processing models. This is completed by utilizing modern text generation models like DialoGPT and T5 Transformers. The model works by using Prompt engineering, Retrieval Augmented Generation (RAG), and fine-tuning techniques on an augmented dataset, text-to-speech engines such as Speech T5 along with Meta's Massively Multilingual Speech (MMS), that can produce naturalistic computer speech in multiple languages with speed that carries the nuances of human communication. The goal is to offer an empathetic, friendly solution that meets the mental health needs of people with varied diversities. Using the unique blend of AI and concern, this research tries to develop an innovative tool for empathic care and support people to be resilient in the face of life's trials, by generating real-world based solutions and responses, including humor as per the user's receptivity to build rapport and indulge the user in engaging conversations and journalling.***

Keywords: Therapeutic Conversational Agent, DialoGPT, Meta's Massively Multilingual Speech, Whisper large v2, Speech T5.

1. Introduction

In the fast-paced modern world of ours, mental health disorders such as stress, anxiety, depression, procrastination, and overthinking are getting more and more popular. The issue of personal counselling and professional assistance for people remains a problem, especially in developing countries, where up to 75% of patients do not receive any assistance or treatment. Studies have shown that such problems can have adverse health consequences and mature into conditions such as Infertility, Heart attacks, Stroke, reduced Stamina, Hypertension, increases susceptibility to Cancer, Obesity due to Stress eating, Irritable Bowel Disease (IBD). Such problems can claim a person's life, starting with destroying his or her motivation, productivity, and, consequently, the overall quality of life. The numbers are shocking – experts calculate that anxiety disorders alone affect over 284 million people worldwide. Counseling is a lifeline, but the problem is that to implement it, you need your own strength and money, as well as enough support from the local system. What is even more alarming is that in the underdeveloped countries, the percentage of the untreated patients is as high as 75%. This indicates a clear lack in emotional support systems.

Comprehension in conversations, is not just understanding – it is understanding back. Existing mental health support systems only respond to a query in a robotic manner and do not include empathy. Instead, the user is bombarded with several questions to understand the complete context, which may sometimes give the complete opposite results to why such systems are created, to support users, and produces very generalized answers, far from personalized support. Our paper exhibits progress in the creation of a conversational AI assistant in the field of mental health support across all ages and diversities. The knowledge retrieval component in the system is integrated with the most recent text generation and summarization models, resulting in short answers to user queries. Particular attention is given to the development of empathic outputs based on a pre-trained dialog model, in which empathy implies a feeling of sympathy and empathy. In addition, this paper examines the integration of text-to-speech (TTS) functionality into the conversational AI system for mental health. The system uses sophisticated TTS models to produce the agent's responses in a natural way. A unique element is allowing the users to make the synthesized voice customizable and changing gender, age.

The proposed system not only gives empathetic and helpful responses to user queries and statements, but also includes humor, if the user is receptive, to lighten the conversations. The review by Ayisire et al. [1], states that Humor has the most potential to eradicate negative emotions in people with

Paper ID: SR24419013403 DOI: https://dx.doi.org/10.21275/SR24419013403

stress and depression. Laughter is the best medicine, Afterall. This personalization seeks to enhance user interaction and confidence. The research looks at the effect of this TTS integration in the situation of mental health care. The system employs a hybrid mixture using DialoGPT and T5 Transformers that forms the basis of our model, text generation. The project also employs Google Translate for translating the text to other language, if needed by the user. We combine Whisper large v2 for Automatic Speech Recognition (ASR) with state-of-the-art text-to-speech engines such as Speech T5 and Meta's Massively Multilingual Speech, to generate synthetic computer speech which retains the subtleties of human interaction. Speech T5 especially performs well at imitating inflections, pauses, and rhythms that demonstrate empathy.

The audio quality of Speech T5 TTS is high, giving our AI the many friendly voices of a supportive friend. Such Existing chatbots, though providing a ray of hope, at present have several limitations that manifest in the inadequacies of their current capabilities. Such constraints comprise lack of personalization and contextual awareness, emotional intelligence, and empathy. Moreover, cultural understanding and availability on a global scale are absent as well as integration in professional care systems. Confidentiality, safety, and ethical concerns continue to be the areas of worry. Lastly, the operationalization of long-term engagement by evaluating the influence of these chatbots is still under development. These are the gaps that need to be addressed for mental health chatbots to show their true potential and offer effective help to persons in need.

2. Literature Survey

Hsu et al. (2023) [2] proposes a transformative approach to empathetic conversational AI systems, utilizing a transformer-based language model integrated with attribute models for both affective and cognitive empathy. The modified model in our research combination with Empathetic Dialogues dataset has modular architecture which enables adjusting language generation dynamically. This feature enhances the empathetic communication of the model with little re-training. DialoGPT and T5, coming together, melding with advanced NLP techniques, data augmentation methods – form a system. Raamkumar et al., (2023) [3], under the IEEE umbrella called 'Transactions on Affective Computing', coverage those and several more empathic conversational AI creation forms by examining the Usage of EMPATHETICDIALOGUES dataset broadening into the research community, whereby the comfort of speech-based unities rose to significance within. Prime in generation our multimodal-having arrived we must establish the stance for researched-oriented development of much more progressive empathetic conducts. Emphatically pointing out then, a proposed method, usage of DialoGPT seen in alliance with T5 and fundamentally driven by notably strong data augmentation strategy in operations combined with more advanced NLP. Karna et al. (2023) [4], observed transformer-based models – such as BERT and RoBERTa - utilized for the purpose of detecting depression in social media platform content. Experimenting with sophisticated GPUs and a cluster of artificial intelligence abundant learning frameworks—covering categories such as LSTM, CNN, and BiLSTM— these models were supplemented with BERT and RoBERTa layers. The methodology underwent improvement: entwining advanced technologies like DialoGPT and T5. With a deliberate integration, the models wound up providing utility as an assembly line of sorts—each singularly focused yet part of an integrated machine—aiming for precision and efficiency in depression detection. Bird and Lotfi, (2023) [5], investigate for people's mental health issues – anxiety or depression to-be-treated – and pondered the efficiency of high-level model chatbots. Chatbots were specifically constructed. The goal: 88. 65% accuracy. Tokens were predicted – calculations revealed 96. 49% and 97. 88% as good predictions. The top 5 did well and the top 10. This hybrid study, combining DialoGPT and T5 technologies, advanced NLP methods and data augmentation – giving chatbots a system of enhancements.

The investigation by S. Alharbi et al. (2021) [6], unveils the new discoveries, noting that transformers, as the most developed deep neural networks, play a crucial part in resilience to noise and accuracy in speech patterns. It exemplifies problems e.g., accent detection, highlighting that large-volume of accumulation data matter and names the way of its future research in this area as multi-channel information processing. Trabelsi et al. (2023) [7], show their comparative study on the use of ASR tools for open sources. Their goals are in accuracy evaluation and inference time test. Research states that Kaldi is very successful on the environments, outperforming Deep Speech, and Kaldi due to its efficiency and robustness should be considered as a model. Through them, model adaptation to new lingo or accent is stressed, which is key in Europe for those who wish to maintain data sovereignty – something businessmen especially will be looking for, regarding the General Data Protection Regulation (GDPR). This research points to the need to do more on those platforms with building custom models for automatic speech recognition that are adapted to the specific enterprise requirements.

Research conducted by Ao et al. (2022) [8], gave life to something new - SpeechT5 framework. Spoken language processing is its specialty, and in its arsenal lies a shared network of decoder and encoder supported by pre-nets as well as post-nets. This ensemble stands, resilient and ready to tackle the tasks of putting speech into text or reverting text into speech. The training data used? Well, an enormous volume of unlabeled data had the privilege of being utilized. The key function it performs is that of converting speech and text into what can be interpreted as having a single meaning. The applications deem this ability quite valuable; they include – recognition of speech, translation of speech into speech and identification of speakers - the latter being one of the multiple tasks. Kim et al. (2021) [9], proposes a new method for text-to-speech generation that uses variational inference and adversarial training to improve expressiveness (VITS), that serves as the base for Meta's Massively Multilingual Speech. Liu et al. (2023) [10] conducted a study focusing on the inadequacies of contemporary speech technology - it does not account for many languages spoken across the world.

To overcome these constraints, an innovative scheme named Massively Multilingual Speech (MMS) was put forth by the authors. They harnessed self-supervised learning on religious texts from dialects numbering over 1,000 to break linguistic barriers. Liberation from traditional methods bore fruit in the form of models useful for deciphering speech, creating speech, and identifying languages. Results showed improvements in the domain of multilingual speech recognition overtaking those seen in prior models without extensive labeled data being required.

Kaur (2023) [11], dives deep into ML techniques' incorporation into mental health analysis. The many uses of ML for analyzing data that is not so structured—think social media, survey answers, even medical records were given a mention. Now if we attempt to illustrate, this bit—still Kaur's words—that weaves together structure and unstructured data may add strain on the gray matter. The Author did not skip over the challenge part. Kaur flagged out the small datasets that are like mere pinpricks. A concern on structured inputs making ML software dependent and oh so less flexible. And to top it off, an acknowledgment was made on its limited accuracy in successful mental health analysis via ML. The system, as a crescendo in the review, asked for something—research of note. It demanded a focal point on methods of a multimodal nature, to improve ML performance in detecting depression. This approach Kaur thought, text, image, video, audio—it has potential, she insisted. It can deal with the messiness that characterizes unstructured patient data. So, is there point to all this? Kaur almost gives you a sense of it—she dared envision ML technology in a world where it is used more effectively in mental health analysis. An advanced system, that is what the authors laid out as the goal.

3. Methods

The goal of this system is to move toward taking a humanistic view of mental health, an approach that is based on a person-centered view. The methodology is continuing finding the roots in the principles of empathy, sympathy and problems solving extends beyond the diagnostic label.

3.1 Text Generation for Conversations

The team plunged into the open-source language models world – BERT, RoBERTa, DistilBERT, and others - to find the perfect match for our mental health chatbot. This is like enormous AI minds, already fed with extensive datasets. This was to save a lot of time and effort in creating a new model from scratch. We trained and fine-tuned DialoGPT, T5 Transformers on our custom dataset for generation of empathetic responses with solutions and an approach to view problems from different perspectives such as addiction point of view, psychological and social lenses. The approach of combining retrieval and generative methods has been demonstrated in the model created by Beredo et al. [12]. The novelty in our approach is of using a custom generated dataset along with concatenation of other publicly available datasets into our dataset after converting them to dialogue pair format, suitable to our format of data, alongside the training of the model to generate empathetic responses with humor, which may be disabled by the user, if not receptive.

BERT, the old text generation model, performed well. It scored decent in the mental health conversations (perplexity of 18.4) and excelled at 81% of the tasks we offered it. But it was not as per the requirements for our system. Its improved version, RoBERTa [12] was more 'comprehensible' (perplexity 15.9), with more tasks done right (85%), and faster responses. Then we also tested out DistilBERT which is a distilled version [14], as the name suggests, smaller, faster, a bit less precise (perplexity 19.2, tasks 78%).

The selection of the correct AI building blocks felt like solving a puzzle. DialoGPT surprised us with the 'understanding' part! On difficult mental health talks, it was the least perplex. (Perplexity 12.7) and excelled most of the tasks we give it (91% completion). DialoGPT uses local attention in every other layer with a window size of 256 tokens, having its architecture like GPT2, but better [15]. Hence, we included DialoGPT, which is a conversational model trained on a unique dataset of Reddit conversations collected over several years. This dataset allows it to generate responses that are informative, engaging, and closely reflect the dynamics of human dialogue, which is exactly what we envision for our model.

3.2 Speech to Text/ Automatic Speech Recognition

We trialed the speech recognizer – Whisper large v2, developed by OpenAI. It was the real-world champion besides Vosk - better at navigating through accents and giving the near accurate transcription, perfect for our task. The application uses the Whisper large language model to transcribe the recorded audio of the user. Whisper is a state-of-the-art speech recognition model that is trained on a massive dataset of around 680K hours of audio and text. The model can transcribe speech with high accuracy, even in noisy environments. Whisper large V2, language pertaining ASR model, OpenAI is the insider of this project, which is used for humanizing mental health app. It is also astounding that whisper has so many parameters (about half billion) which means that it abounds in enhancing speech recognition tasks.

The project pivots on the Whisper large V2 Automatic Speech Recognition (ASR) model from OpenAI to translate the audio. On one hand, the Whisper platform may not be directly trained for the specific task of mental health data processing, but on the flip side, its amazing variety of multilingual capacities and audio characteristics are quite encouraging and could be a starting point for further research in mental health data processing.

Comparing the transcript of each model against the original audio recording was crucial here and careful observation is needed to determine how accurately a model can convey the naturalness of human speech. We considered aspects such as accuracy of transcription, management of the emotional tone, prevention of hesitation, and other strains that can affect the mental health conversations. Rather than quantitative approach, this work focuses on qualitative one, the major result of which reveals the strong and weak points of the whisper large V2, the baseline model in the framework of our application.

We evaluated the performance of ASR system using Word Error Rate (WER) and accuracy. WER provides a detailed measure of errors made by the system, while accuracy shows the overall percentage of correctly transcribed words.

$$\text{Word Error Rate (WER)} = \frac{(\text{Substitutions} + \text{Insertions} + \text{Deletions})}{\text{No. of words (Reference)}}$$

Where, Substitutions refer to the number of words that have been transcribed incorrectly by the ASR system as compared to the original/ reference transcription. For instance, if the reference transcription is “The feline took a seat on the desk” and the ASR transcription is “The female took a seat on desk”,
The WER would be calculated as, Substitutions = 1 (‘feline’ as ‘female’), Insertions = 0 (No new word inserted), Deletions = 1 (‘the’ missing in ASR Transcription), No. of words in reference transcription = 8. Hence the WER would be 2/8 i.e., 25%.

Similarly, for the calculation of accuracy, we used a predefined formula in line with similar research, with the same meanings for every term as for WER, which is as follows:

$$\text{Accuracy} = \frac{(\text{No. of words} - \text{Substitutions} - \text{Insertions} - \text{Deletions})}{\text{No. of words in reference} * 100}$$

Table 1: Comparison of ASR models

Model	Word Error Rate (WER)	Accuracy (%)
Vosk API	9.7%	90.3
Whisper v2 (Large)	8.5%	91.5

3.3 Text to Speech

Adding multi-modality to such conversations can give a personalized touch and help the user build trust, for better usability. The team conducted rigorous testing and trial of some of the most popular Text-to-Speech models such as Tacotron 2 [15], a text-to-speech system known for generating naturalistic human-like speech with natural inflections and pauses. It uses a neural network architecture with an attention mechanism and a post-processing vocoder to create realistic speech directly from text. Another model, VITS, is a text-to-speech system known for its efficiency and ability to generate high-quality speech. It uses fewer parameters than similar models. It utilizes a neural network architecture based on variational inference and combines the strengths from autoregressive and GAN-based text-to-speech models.

Speech T5 model is a flexible model which perfectly handles different speech tasks such as speech recognition, translation, and text-to-speech synthesis. Based on the T5 Transformer architecture, it is a master in comprehending and manipulating spoken language. It provides a coherent method for text-to-speech, producing audio waveforms from the text without intermediary representations, which leads to better quality and simplified pipelines. Additionally, it permits manipulation of speech features like pitch, speed, and emphasis allowing for the generation of sophisticated, more subtle, and natural-sounding artificial voices. It is quick and effective and caught us unaware due to its multitasking capabilities. The audio was natural with low disturbance (MOS 4.2), and was easy to

For supporting Multilingual Conversations, we employed Meta’s Massively Multilingual Speech, which is a model developed upon VITS, by training it over religious text reader recordings in more than 1000 languages, and using it to convert the translated text (from English to the required language) to Speech.

Lastly, Matcha-TTS [16], a very novel text-to-speech (TTS) framework was tested, that generates high-quality speech in less steps by utilizing a novel training technique called optimal-transport conditional flow matching. It was taken into consideration due its phenomenal generation speed, less memory footprint, and highly clear and natural outputs. But it offered lesser customization as fine-tuning it was difficult, being a very recent model. It serves as the most ideal option for mobile applications requiring the production of high-quality, human-like voices, having the least memory footprint while offering unparalleled speeds. If trained and fine-tuned properly, using the appropriate resources, it may become customizable. Thus, as is evident from the figures, was why we chose Speech T5 for our model.

The Mean Opinion Score was calculated as usual, by calculating the mean of ratings given by the listeners out of 5. The throughput is calculated by taking the number of samples processed and dividing it by the total time taken.

$$\text{Throughput} = \frac{\text{Number of Samples Processed}}{\text{Time taken for processing}}$$

i.e., if the number of samples processed by a TTS system is 5 in 14 seconds, the Throughput is 5/14, i.e., 0.36 samples/second.

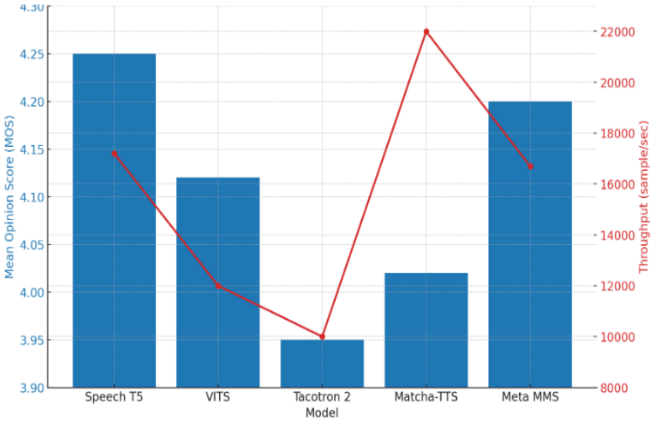

Figure 1: Comparison of TTS Models by MOS and Throughput

3.4 Dataset

The dataset has been created using PersonaChat data by Meta, cleaned dataset from Reddit containing 7,650 records, Mental *Health Corpus containing text and records related to anxiety, depression, and mental health issues with 27,972 records and a dataset named Mental health chatbot created by Mark Daniel Lampa with around 100 dialogue pairs.*

Table 2: Dataset details with Evaluation metrics

Metric	Value
Quantity of dialogues	10,356
Aggregate of dialogue turns	82,848
Peak amount of turns per dialogue	12
Sum of evaluations	10,000
Mean amount of turns per dialogue	8
Minimum amount of turns per dialogue	4
Rate of accurate evaluations	9,956
Rate of inaccurate evaluations	44
Average likeness score	0.69

Apart from Dialogue and Mental Health Datasets, we have also trained our model on Short Jokes Dataset, available on Kaggle, with 2,31,657 entries, which were significantly reduced after cleaning and preprocessing. The dataset for text generation was preprocessed, filtered, and cleaned to include only those entries relevant to our purpose, leaving the quantity to 10,356 entries with around 80 K turns of dialogues.

3.5 Integration of Models

The respective codes of the ASR model for Speech-to-text conversion i.e., Whisper Large v2 model, Text/Dialogue Generation and Summarization i.e., T5 transformers, DialoGPT, Text-to-Speech using Speech T5 along with Translation modules were all integrated together as one model, to ensure the coherent performance of our model. An Application Programming Interface (API) was created for the whole code, and then integrated to the backend of our website application.

3.6 Testing

There were rigorous manual testing rounds, incorporating feedbacks from users and mentors to improve the usability and performance of the model. Additionally, ChatGPT was employed to create test cases (human-like) for testing. We also tested the model for gauging its multilingual performance.

4. Results

This paper dives into the main natural language processing tasks such as text generation, summarization, translation, automatic speech recognition (ASR), and text to speech (TTS) synthesis. We used a powerful ensemble of cutting-edge methods for the purpose of dealing with every challenge in a correct way.

For text generation, we relied on DialoGPT, and T5 Transformers which were beneficial to us because of their ability to exhibit contextually appropriate and diversified outputs. Google Translate was our resort to a translation task, proving to be efficient in converting the text across various languages. In the ASR realm, Whisper Large-v2 model became our main tool, it was precise in terms of input speech translation into text even in difficult acoustic environments. As for TTS synthesis, Speech T5 showed superior performance in creating high-quality synthetic speech in English, and Meta's Massively Multilingual Speech model appeared to be professional in many languages. However, Meta’s MMS does not generate as natural results as the former, but it must be noted that the number of languages supported in the latter model is beyond the scope of any other TTS model, at present.

5. Conclusion

In this work, we proposed a novel multimodal conversational agent capable of understanding user queries and generating empathetic, human-like responses for mental health support and inculcating humor in conversations, if permitted by the user. By leveraging large language models like T5, DialoGPT and incorporating Speech recognition with whisper large v2 by OpenAI and Text-to-Speech with Speech T5 and Meta’s Massively Multilingual Speech, the model can engage in natural and empathetic two-way conversations via text and audio. If the user is receptive towards humor, it also adds humor in its responses, to make the conversations natural and light-hearted. The whole model largely works through Retrieval Augmented Generation (RAG) of responses from the dataset, if similar queries are already present, otherwise, generates new responses altogether, alongside prompt engineering, fine-tuning on our custom created dataset, made by Web Scraping and Data Augmentation. Our experiments showed that the agent was able to understand intentions behind queries, have empathetic discussions and lighten moods by generating funny replies when users were receptive. While still a work-in-progress, we believe conversational AI systems like ours can play a big role in making mental healthcare more accessible worldwide, whilst significantly reducing stress and positively impacting anger. Future work will focus on expanding the agent's knowledge base, improving personalization, and conducting user studies to evaluate real-world impact.

6. Future Scope

Further studies may optimize the system including the use of art to reduce anxiety by honing in on cultural nuances in linguistics and deepening the emotional understandability, comprehension of context and greater awareness. There should also be more and more inclusion of languages across the world, and support for all the communities. The inclusion of predictive modeling can come in handy for the proactive anticipation of the users’ needs and emotional states from behavioral patterns and the real-time data. Thus, the support will be personalized and effective. There should be advanced channel for making the depressed and affected individuals access Expert Support and Support Organizations. Renewing constant learning mechanisms will provide the system with the ability to learn from user interactions keeping previous knowledge intact consequently, ensuring adaptation. Developing in-depth user profiles would take dynamic personalization to a new level by adjusting it in real-time to likes and dislikes, all while preserving users' information. These innovations are the lever which will considerably drive up the quality of human interaction and provide personalized assistance and can even intensify the engagement at workplaces and universities.

References

[1] Ayisire OE, Babalola F, Aladum B, Oyeleye-Adegbite OC, Urhi A, Kilanko A, Agbor C, Adaralegbe N, Kaur G, Eze-Njoku C, Soomro F, Eche VC, Popoola HA, Anugwom GO. A Comprehensive Review on the Effects of Humor in Patients With Depression. Cureus. 2022 Sep 17;14(9):e29263. doi: 10.7759/cureus.29263. PMID: 36262951; PMCID: PMC9576124.

[2] J. -H. Hsu, J. Chang, M. -H. Kuo and C. -H. Wu, "Empathetic Response Generation Based on Plug-and-Play Mechanism with Empathy Perturbation," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 2032-2042, 2023, doi: 10.1109/TASLP.2023.3277274.

[3] A. S. Raamkumar and Y. Yang, "Empathetic Conversational Systems: A Review of Current Advances, Gaps, and Opportunities," in IEEE Transactions on Affective Computing, vol. 14, no. 4, pp. 2722-2739, 1 Oct.-Dec. 2023, doi: 10.1109/TAFFC.2022.3226693.

[4] P. Karna, S. K. Keshari, A. Kumar Mandal and B. Chakraborty, "BERT-Driven Deep Learning Approach for Depression Detection in Social Media Posts," 2023 1st International Conference on Optimization Techniques for Learning (ICOTL), Bengaluru, India, 2023, pp. 1-6, doi: 10.1109/ICOTL59758.2023.10435285.

[5] Bird, J.J., & Lotfi, A. (2023). Generative Transformer Chatbots for Mental Health Support: A Study on Depression and Anxiety. Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments.

[6] S. Alharbi et al., "Automatic Speech Recognition: Systematic Literature Review," in IEEE Access, vol. 9, pp. 131858-131876, 2021, doi: 10.1109/ACCESS.2021.3112535

[7] Trabelsi, A., Warichet, S., Aajaoun, Y., & Soussilane, S. (2022). Evaluation of the efficiency of state-of-the-art Speech Recognition engines. *Procedia Computer Science*, [volume], [pages]. https://doi.org/10.1016/j.procs.2022.09.534

[8] Ao, J., Wang, R., Zhou, L., Wang, C., Ren, S., Wu, Y., Liu, S., Ko, T., Li, Q., Zhang, Y., Wei, Z., Qian, Y., Li, J., & Wei, F. (2021). SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing. *ArXiv*. /abs/2110.07205

[9] Kim, J., Kong, J., & Son, J. (2021). Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. *ArXiv*. /abs/2106.06103

[10] Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., Baevski, A., Adi, Y., Zhang, X., Hsu, W., Conneau, A., & Auli, M. (2023). Scaling Speech Technology to 1,000+ Languages. *ArXiv*. /abs/2305.13516

[11] V. Kaur and K. Gupta, "A Brief Review of Machine Learning Methods used in Mental Health Research," 2023 International Conference on Artificial Intelligence and Applications (ICAIA) Alliance Technology Conference (ATCON-1), Bangalore, India, 2023, pp. 1-6, doi: 10.1109/ICAIA57370.2023.10169520. keywords: {Deep learning;Technological innovation;Statistical analysis;Education;Decision making;Mental health;Medical services;Machine Learning;depression;anxiety;mental health;innovation and health;healthcare;Artificial Intelligence}.

[12] J. L. Beredo and E. C. Ong, "A Hybrid Response Generation Model for an Empathetic Conversational Agent," 2022 International Conference on Asian Language Processing (IALP), Singapore, Singapore, 2022, pp. 300-305, doi: 10.1109/IALP57159.2022.9961311. keywords: {Training;Adaptation models;Oral communication; Mental health;Medical services; Chatbots;Hybrid power systems; hybrid conversational model; dialogue response generation; empathetic responses}

[13] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv*. /abs/1907.11692

[14] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. ArXiv. /abs/1910.01108

[15] Zhang, Y., Sun, S., Galley, M., Chen, Y. -C., Brockett, C., Gao, X., Gao, J., Liu, J., & Dolan, B. (2020). DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation. arXiv:1911.00536 [cs.CL]. Retrieved from https://doi.org/10.48550/arXiv.1911.00536

[16] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Saurous, R. A., Agiomyrgiannakis, Y., & Wu, Y. (2017). Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. *ArXiv*. /abs/1712.05884

[17] S. Mehta, R. Tu, J. Beskow, É. Székely and G. E. Henter, "Matcha-TTS: A Fast TTS Architecture with Conditional Flow Matching," ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, Republic of, 2024, pp. 11341-11345, doi: 10.1109/ICASSP48485.2024.10448291. keywords: {Signal processing; Probabilistic logic; Acoustics; Decoding; Speech processing; Diffusion models; flow matching; speech synthesis; text-to-speech; acoustic modelling

Author Profiles

Smit D. Gaikwad, a final year student of B.Tech Computer Engineering majoring in Artificial Intelligence from Ajeenkya D.Y. Patil University, is a lover of Cloud Computing and Python, with experience as a Python Developer and Analyst at Tata Communications Ltd. They are those who aim to employ what they learn in academics and can work in the real world or the tech industry. Smit is an apt and keen learner. He is also equally a team player and the critical thinker. He is an observant person whose attention to detail never goes unnoticed. They are impactful at keeping to their word and deliver impeccable solutions in designated time frame.

Aishwarya B. Iyer, a final-year B.Tech-Computer Engineering student, with a major in Artificial Intelligence, from Ajeenkya D.Y. Patil University with experience as a software engineer intern at KPIT and TATA Communications Ltd. Her interests cover a wide range of therefore Machine Learning, Data Science, SQL, Web Development, among others, and she also is familiar with Python, C++, R and Java. Aishwarya's interests do not wane from the technical aspects. She also follows interests like creative sketching and has a penchant in statistics.

Krishna S. Talluri, a final year student of B.Tech Computer Engineering from Ajeenkya D.Y. Patil University and specializes in AI. Not only with a command over pandas and Microsoft Excel, they are dedicated in bringing out the best code that is fully optimized and scalable. They will always seek to be abreast with the latest technology insights and industry best practices.