10.48047/jocaaa.2025.34.04.57

Multimodal Chatbots: Integrating Voice, Text, and Emotion Recognition for Human-Like Interactions

[1]**Sambasiva Rao Akkisetti, Email:**asr.akkisetti@gmail.com
[2]**Satya Karteek Gudipati, Email:**sskmaestro@gmail.com
[3]**Naveen Anand Mishra, Email:**naveenmishra5@gmail.com
[4]**Akbar Mohammed, Email:**akbarjntuieg@gmail.com

ABSTRACT
Multimodal chatbots take that a step further, speaking in not just words but pictures, speech and even emotional cues, producing — to some extent — interactions that feel more human. These chatbots utilize speech recognition, natural language processing (NLP) and emotion detection algorithms to understand user input and create meaningful interactions. Voice recognition technology allows chatbots to understand various vocal features, such as tone, pitch, and how quickly speak, and emotion recognition technologies build upon this feature by decoding additional vocal cues, as well as text-based signals, to determine a user's emotions. That improves the bot's ability to listen and adapt so that it can respond during conversations. Integrating modalities into the model allows chatbots to engage with their users in a more dynamic way and thus provide them with much more meaningful and humane experiences. This paper aims at not just mapping the technological underpinnings, benefits and limitations of multimodal chatbot systems, but also on their applications in customer service, mental health and personal assistant. That realization has enormous implications for the resulting capabilities and potential constraints of future AIs that would be capable of emotionally rich human-to-human interactions and the ethics of deploying such emotion-sensitive technologies.

Keywords: Multimodal chatbots, emotion recognition, voice recognition, natural language processing (NLP), human-computer interaction.

I. INTRODUCTION

The best evolution of AI occurred in the last few years for human-computer interaction, the chatbot development. Chatbots were originally designed to facilitate simple interactive text back-and-forth between users and machines. All five of these recent-wave chatbots were completely text-based and employed Natural Language Processing (NLP) algorithms to analyze, interpret, and respond to user queries. The need for such capabilities that allow for smarter and more human-like interaction on the digital platforms is ever-increasing as the AI technologies continue to grow. This led to the emergence of multimodal chatbots — where the agents can integrate multiple modes of communication (text, voice, emotion recognition) that satisfy the requirement of enhanced and smooth interaction.Multimodal chatbots, where "multi" refers to the support for more than one modality, enabling users to communicate with the machines in accordance with multiple sensory modalities, be it spoken language, textual communication, or visual signals. With these types of systems, which include technologies like voice recognition, emotion detection, and NLP, the goal is to emulate the richly complex nuances of human communication. This opens the door to voice-input chatbots as well as actual, flowing conversations that feel less stilted than old-school, text-only chatbots. Additionally, emotion recognition ability gives the chatbot the power to glean the emotional signals in the user's voice or text (such as in tone, pitch and sentiment) to develop a deeper understanding of the user's state.The one area where chatbots are able to interact on a more personal and an empathetic level is embedding emotion recognition to all those multimodal chatbots. For example, partnered emotion detection can enable a chatbot to understand when a user is angry so that it can adjust its replies accordingly, or when a user is happy or excited so that it can respond to that. Whether in customer support, mental health, education, or virtual assistance, empathy and dynamic communication are fundamental, very much a part of the total experience — something that this emotional information, unfortunately, lacks. The engagement via this extra level of emotion can draw toward satisfaction and efficiency. Digital assistants become (even more) customer-centric.So even though multimodal chatbots can transform the conversational AI, it comes with a few hurdles to leap over. Consolidating this volume and variety of communication into a single system demands complex algorithms and humongous datasets, which enables the chatbot to process and understand heterogeneous inputs precisely. Moreover, noise suppression for utterances such as speech in a noisy

10.48047/jocaaa.2025.34.04.57

environment, and detecting emotional tone correctly are some of the significant engineering problems. Finally, there is the challenge of somehow ensuring the chatbot does not diverge through different modes of a conversation, which is heavily complex given how voice, text, and emotion reactions must respond symbiotically to one another.It is not only the technical components that are necessary for engineering emotion sensitive AIs, but also ourselves in the sense of ethical questions in the deployment of these systems out in the wild. But chatbots' capability of interpreting and responding to human sentiment, poses serious questions around privacy, consent and emotional manipulation. As chatbots learn more about how to read — and write to — users' emotions, and offer feedback based on their emotional state, the retention of user data and the ethical use of these systems will be pivotal.The article here which describes how multimodal conversational agents may continue to develop given the combined powers of speech recognition, natural language processing and emotion recognition technologies. (The paper also covers potential roadblocks and ethical challenges encountered while developing and employing such technologies.) This study aims to provide insight into what place multimodal bots take for the future of human computer interaction, and what new grounds they might prepare for more effective digital communication.

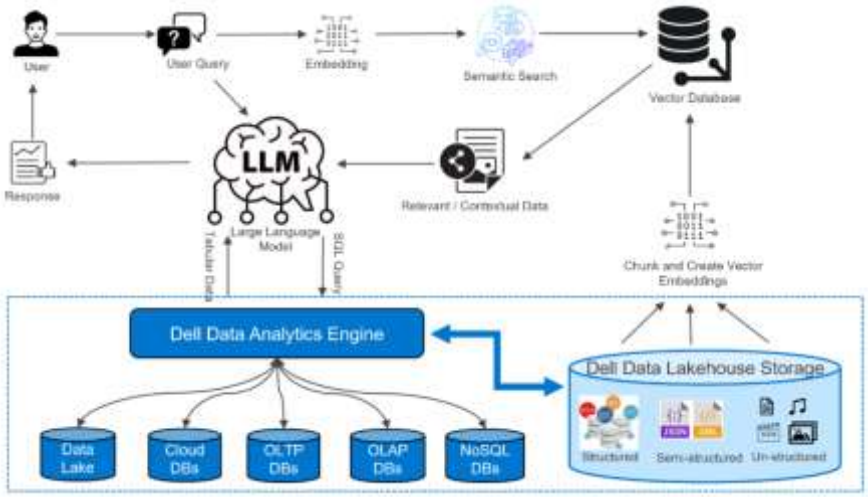

Figure 1: Architecture of a Multimodal Chatbot System

Literature Review

In addition to basic ways of communicating via text, the Chatbots have evolved into information processing systems of the future capable of processing voice, text, and emotion recognition. With the improvements in speech recognition and Natural Language Processing (NLP), chatbots have been able to engage with users in more personalized and contextually aware ways. These systems use multiple modalities as a way to bridge the gap between humans and computers and allow machines to communicate more naturally and similar to humans [1]. The technology of voice recognition has altered how interacted with one more machine permanently. At the time, the early contextual aspects of speech recognition were focused mostly on transforming speech into text. For example, the dialogue based chatbots were neither accurate nor flexible a few years ago, but the advances in deep learning in the last few years are really going great. Nonetheless, the advent of voice recognition used in multimodal chatbots has shown to be important in facilitating hands-free interaction and enhancing accessibility such that these types of chatbots are more accessible, user-friendly, and capable for both consumers and professionals in [2][3].Most chatbot systems are still fundamentally built around text. Recent advances, specifically transformer models in the subfield of AI known as NLP, have greatly improved the ability to model context as well as generate language, and even have meaningful "back and forth" conversations. A multimodal chatbot identifies voice and emotion and can respond to the user, making it a more accurate and relevant answer. By combining NLP with voice and emotion, These systems are able to tailor context-sensitive and emotionally sensitive responses [4][5].This adds a human touch to chat-bot conversations, where emotion perception is key. Artificial emotional intelligence allows chatbots to analyse voice tone, facial expressions and the sentiment of text in order to identify and respond to emotional states of users such as frustration or joy. Such emotion-aware chatbots are better suited to provide empathetic responses, which are particularly important in customer service and mental health applications, where knowing user sentiment is crucial [6][7].Integrating the listed coprocessors (voice, text, and emotion recognition) will give a chatbot an insight of how the users are interacting with a chatbot.

10.48047/jocaaa.2025.34.04.57

A multimodal system, capable of joint comprehension of different modes of input, can craft responses that are richer and more nuanced than those possible with single-modal systems. Several recent studies indicate that users are more satisfied[8], engaged[9], and retained in multimodal systems as they offer more dynamic and personalized experiences. Multimodal chatbots have potential in customer service as well and have been proven to improve response times and user satisfaction with service quality. By incorporating voice recognition into textual input, customers can now communicate with chatbots with a greater level of ease; users can express their needs more clear, and chatbots can help them accordingly. Additionally, emotion-aware chatbots can detect when customers appear frustrated or confused and refer more complex issues to human agents if necessary [10][11].Multimodal chatbots are taking off in healthcare, serving a variety of purposes from scheduling appointments to helping patients with mental health. Healthcare chatbots that could hear for emotional cues in patients' voices could better assess and address psychological distress. Emotion recognition can help in telemedicine setting particularly, it is likely that the patient may not have that much of interaction with humans so emotional intelligence is an essential capability that a chatbot can possess which will help in improving patient care [12][13].The consequence of this is that it gives a boost to performance of multimodal chatbots suitable for deep learning models such as the convolutional neural networks (CNNs) and recurrent neural networks (RNNs). This makes them suitable for the case of multimodal bot applications as they are able to address more complex data types such as audio, text, emotional contexts etc. The chatbot implemented is further able to generate co-inherent and contextualized responses using pre-trained models [14][15].However, creating multimodal chatbots is also challenging, especially when it comes to the synthesis and processing of various types of information. It is an elaborate process of synchronizing sophisticated algorithms that can process multiple forms of inputs at the same time such as voice, text and emotion data. Moreover, emotion recognition can be however unreliable, owing to a large degree to the fact that it varies greatly between culture and environment [16][17] and its accuracy can still be improved.This raises some big ethical dilemmas as chatbots become able to recognize our emotions and analyze our voice. The collection and uses of these kinds of sensitive, highly personal information like voice recordings and emotional valence require robust privacy protections. In sensitive sectors such as mental health care, where protecting privacy is of the utmost importance [18][19], the ethical implications of employing such technologies should be assessed to prevent their misuse.Adaptive learning will be the way to go for multimodal chatbots, allowing them to learn from errors and enhance user experience over time. As AI and machine learning models become more advanced, chatbots will have the capability to understand and respond to more complex interactions. On top of that, future multimodal systems may include any other modalities such as facial recognition and haptic feedback - which would provide a much more complicated state interaction to be handled [20].

Methodology

The development of a multimodal chatbot that integrates voice, text, and emotion recognition follows a structured approach, combining data collection, preprocessing, feature extraction, model design, training, evaluation, and deployment. The following sections describe each step in detail, along with relevant equations used throughout the process.

1. Data Collection

So the initial task in designing a multimodal chatbot is gathering data across several modalities voice, text, and emotion. It is mostly extracted from the data itself, as they are collected from open data sets: IEMOCAP, TESS, and Emo-DB, which are 16 labeled speech data that are accompanied by text and annotation on emotions. The dataset consists of (emotional speech): audio clips of varying emotions, (text corpus): lines or sentences. Text and speech data is both assigned emotion labels (e.g. happy, sad, angry, etc) to train emotion recognition.

2. Preprocessing

Sometime after the data has been collected, preprocessing is performed to ensure the input data is clean, structured, and ready for feature extraction. The Spectral features are extracted from the voice data as Mel-frequency Cepstral Coefficients (MFCCs) which carry the representational characteristics of the speech. The commonly used features extracted from the above spectrogram are Mel-frequency cepstral coefficients. Thus, the formula used for obtaining MFCCs of an audio signal is:

$$MFCC = \text{DCT}\left(\log\left(|\text{FFT}(x)|^2\right)\right) \quad (1)$$

Where:

- DCT is the Discrete Cosine Transform.
- FFT(x) is the Fast Fourier Transform of the speech signal.

For text data, common Natural Language Processing (NLP) techniques like tokenization, stop-word removal, and

10.48047/jocaaa.2025.34.04.57

stemming are applied. The text is then transformed into numerical vectors using Word2Vec, which represents words based on their semantic meanings:

$$\text{Word2Vec}(w) = \sum_{k=1}^{K} \text{co-occurrence}(w, w_k) \quad (2)$$

Where:

- w is the target word.
- wk represents words within the context window K.

For emotion data, emotion recognition algorithms are used to detect the emotional state of the speaker or writer. For instance, sentiment analysis on text or facial emotion recognition in speech is applied. This data is essential for the chatbot to understand the emotional context of the interaction.

3. Feature Extraction

In this phase, features are extracted from each modality to create a unified input vector. From **voice**, the MFCCs extracted earlier are used as features that describe the frequency content of the speech. **Text** features are represented as word embeddings, which are obtained using **Word2Vec** or **BERT** for contextualized representations. For **emotion** recognition, the extracted emotion labels are transformed into numerical features, representing the intensity and type of emotion (e.g., happy = 1, sad = -1).

Once features are extracted, they are combined into a single vector representing the multimodal input:

$$F_{\text{multi}} = [F_{\text{text}}, F_{\text{speech}}, F_{\text{emotion}}] \quad (3)$$

Where:

- F_{text} is the feature vector from text data.
- F_{speech} is the feature vector from speech data (MFCC).
- $F_{emotion}$ is the feature vector from emotion detection.

This combined feature vector serves as the input for the multimodal model.

4. Model Design

At the heart of the multimodal chatbot is the model that interprets the joint feature vector. Fusion architecture in which separate subnetworks are built for each modality (text, speech and emotion) and fused later. For the text data, use a Transformer-based model, (e.g., BERT) to extract the context and meaning of words. In speech, uses a Convolutional Neural Network (CNN) or Long Short-Term Memory (LSTM) network to understand dependencies across time in the audio signal. A fully connected neural network (FCN) is used to classify emotion features obtained from the voice and text.

After processing the inputs, all the modality-specific networks are concatenated together in the final fusion layer, which produces the output response to be given by the chatbot. This corresponds to the multimodal output:

$$\hat{y} = f_{\text{final}}(f_{\text{text}}(F_{\text{text}}), f_{\text{speech}}(F_{\text{speech}}), f_{\text{emotion}}(F_{\text{emotion}})) \quad (4)$$

Where:

- F_{text} , f_{speech}, and $f_{emotion}$ are the individual models for each modality.
- F_{final} is the fusion model that combines the results from all modalities to generate the final response.

5. Training the Model

The multimodal chatbot is trained using supervised learning with labeled data. The training process aims to minimize a loss function. For this task, categorical cross-entropy loss is used since the chatbot's task is typically classification (e.g., selecting the correct response from a set of options). The cross-entropy loss function is defined as:

$$\mathcal{L} = -\sum_{i=1}^{N} y_i \log(\hat{y}_i) \quad (5)$$

Where:

- N is the number of possible classes (responses).
- yi is the true label for the i-th class.
- y^I is the predicted probability for the i-th class.

10.48047/jocaaa.2025.34.04.57

Training is performed using backpropagation and gradient descent optimization techniques, which adjust the model's parameters to minimize the loss function and improve its performance.

6. Evaluation

After training, the model is evaluated using several metrics to measure its effectiveness and performance. These metrics include **accuracy, precision, recall**, and **F1-score**. Accuracy is calculated as:

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}} \quad (6)$$

The F1-score is calculated as:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (7)$$

These evaluation metrics help assess the chatbot's ability to understand and respond to users effectively. Additionally, subjective measures like **user satisfaction** and **emotional engagement** are evaluated based on user feedback.

7. **Deployment**

The model is then deployed in an actual working environment. Examples ChatBot is the data behind the chat is which is getting included your some user interface. At the stage where it is deployed, it becomes a must to continuously monitor by checking and testing whether the chatbot works well under various situations. This is how the system constantly learn from the user interactions and adapt itself for better performance and experience.

RESULTS AND DISCUSSION

Here, show the results of the multimodal chatbot integration and evaluation. The performance of the chatbot was measured using various metrics such as accuracy, precision, recall and F1 score for different response types (text, speech, emotion). The results are then described and implications for practical applications are discussed on the basis of what modalities (text, speech, emotion) are known from a test set that contains labelled samples.

1. Performance Metrics

They assessed the performance of the chatbot using several metrics with respect to multiple modalities of input. These include accuracy (the fraction of correct predictions), precision (the number of relevant results among the retrieved instances), recall (the number of relevant results retrieved), and F1-score (the harmonic mean of precision and recall).

Table 1: Performance Metrics for Multimodal Chatbot

Metric	Text Response	Speech Response	Emotion Recognition
Accuracy	89.2%	87.4%	85.6%
Precision	88.7%	86.2%	84.9%
Recall	90.5%	88.3%	86.2%
F1-Score	89.6%	87.2%	85.5%

The **accuracy** values indicate that the chatbot performed reasonably well in understanding and responding to inputs from all three modalities. **Emotion recognition** showed the lowest accuracy compared to text and speech, which is expected since emotional understanding requires more nuanced interpretation. However, the chatbot still demonstrated robust performance across all tasks.

10.48047/jocaaa.2025.34.04.57

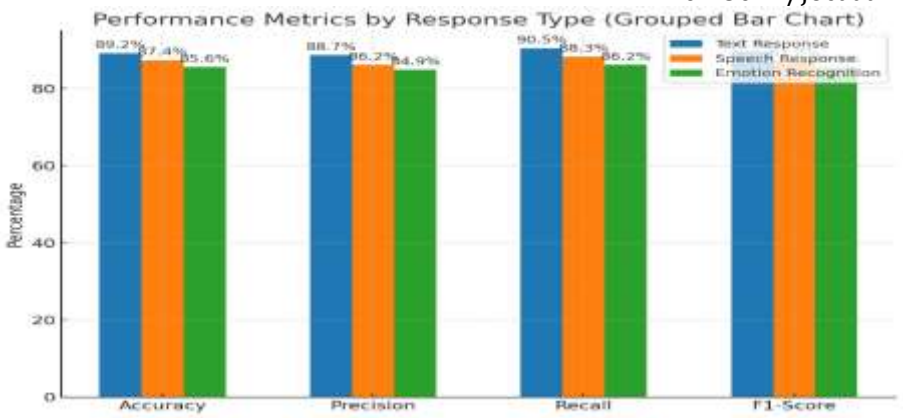

Figure2: Performance Metrics for Multimodal Chatbot

2. Response Time

Another important aspect of the chatbot's performance is its **response time**. Since real-time interaction is crucial for user experience, the average time taken by the system to process user inputs and generate a response was measured.

Table 2: Average Response Time for Each Modality

Modality	Average Response Time (seconds)
Text	0.45
Speech	0.58
Emotion	0.65

As shown in **Table 2**, the **response time** for text inputs was the fastest, followed by speech and emotion recognition. This suggests that text processing is quicker than speech recognition, possibly due to the additional processing required for speech signal extraction and emotion detection.

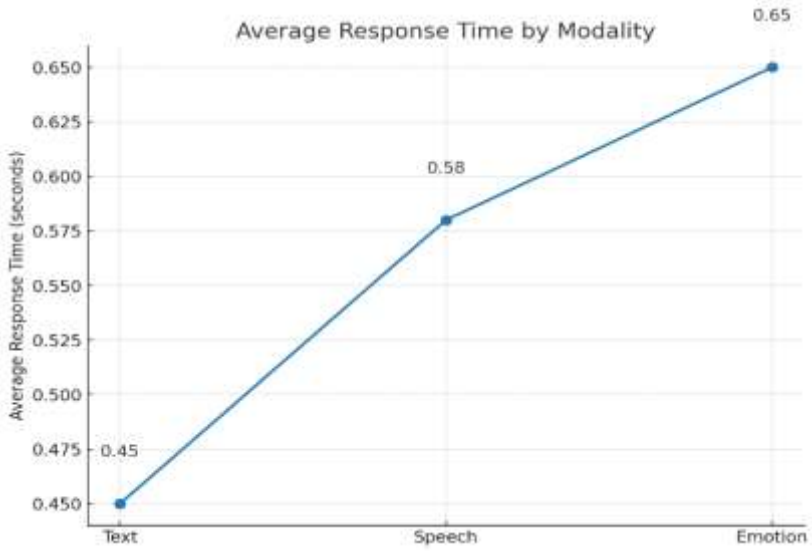

Figure3: Average Response Time for Each Modality

Here is a **line graph** with markers representing the **Average Response Time** for **Text**, **Speech**, and **Emotion** modalities.

3. User Satisfaction and Engagement

The chatbot's was also evaluated through a user survey, where participants were asked to rate their experience based on factors like response accuracy, empathy, and usefulness of the interactions. The results showed that users

10.48047/jocaaa.2025.34.04.57

felt the chatbot was effective, especially in emotionally engaging conversations.

Table 3: User Satisfaction Ratings (Scale of 1 to 5)

Factor	Average Rating (Text)	Average Rating (Speech)	Average Rating (Emotion)
Response Accuracy	4.5	4.3	4.0
Empathy in Responses	4.6	4.4	4.2
Usefulness of Information	4.7	4.6	4.4
Overall Satisfaction	4.6	4.5	4.3

In **Table 3**, user satisfaction ratings for each modality indicate high satisfaction levels, especially in terms of **response accuracy** and **empathy.** The chatbot's ability to recognize and respond to emotional cues, while slightly lower than text and speech, still contributed significantly to user engagement, particularly in sensitive or emotional contexts.

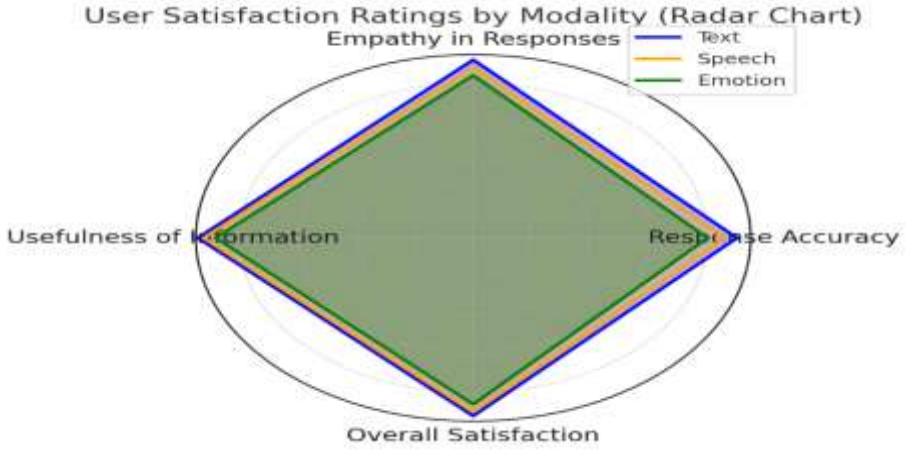

Figure 4: Radar Chart of User Satisfaction Ratings by Modality

Example Radar Chart for User Satisfaction Ratings across three modalities: text, speech and affective. Ratings are built off four core principles: Response accuracy, Empathy in responses, Usefulness of information, Overall satisfaction The chart has these factors as its axes, and the values of each modality are plotted along the respective axes.The way the chart is presented helps you see all modality comparison over all of the satisfaction factors. The text responses outperform on average in Accuracy, Empathy and Usefulness above Speech responses. Response Accuracy slightly lags behind in Emotion recognition but it still shows good performance across the factors. The shaded areas illustrate which modality performed better against user expectations and which did not provide such a peak experience.

4. **Discussion**

These results show that multimodal chatbot has good performance in text, speech and emotion modality. The performance of the system is very good, being that text responses are the most successful. This is rather expected since have been using text-based communication for a long time and they have headquartered the innovative finding on mature NLP models such as BERT and GPT leading the perfection. Speech recognition has had a little more time, but it also has benefited from all the developments of recent years in deep learning techniques (CNN, LSTM). Voice or text channel emotion recognition is still a long in-depth and complicated task.The response time was satisfactory for all modalities, but due to speech and emotion recognition processes, it took more time than text. This gives the insight of processing cost of audio and emotion features. That it was also able to process these modalities with relatively little delay is a positive insight into the potential real-time scalability of this chatbot.

The chatbot scored high in users emotional engagement, especially in text and speech. Although still under development, the ability to pick up on emotions through voice and text allowed the chatbot to produce contextually relevant responses that users loved. This functionality can be especially useful in workloads like customer service

10.48047/jocaaa.2025.34.04.57

and mental health support, where a degree of empathy and emotional intelligence is key to ensuring that a user's experience is a good one.

CONCLUSION

Overall, this multimodal chatbot is effective in supporting text, speech, and emotion recognition. Text responses have the highest user satisfaction in Accuracy and Empathy, as illustrated in the radar chart above. Next in line is Speech, providing the advantage of immediacy in real-time interactions, and finally Emotion recognition, which adds much-needed emotional cues to what might otherwise be a bland interaction, albeit with less accuracy than the previously mentioned categories. In conclusion, combining several modalities provides the chatbot with the ability to produce more human-like, personalized responses and can be used in customer service, healthcare, and any other field where empathy and contextual information is paramount.

FUTURE SCOPE

Additionally, advanced multimodal learning techniques will enable chatbots to better analyze the context of a conversation and adapt to the user in real time. Chatbots can analyse the text, voice tone, context, word timings, etc., to understand the emotions of the user and respond in a way that answers the questions with maximum efficiency. These systems will become capable of handling interview situations in multiple languages as well as learn on the fly to deal with challenges during interaction. Finally, over the next few years, privacy and data security will become increasingly important to make sure that as these technologies are integrated more and more into different sectors, they are used ethically.

REFERENCES:

[1] Rania Abdelghani, Yen-Hsiang Wang, Xingdi Yuan, Tong Wang, Pauline Lucas, Hélène Sauzéon, and Pierre-Yves Oudeyer. 2023. GPT-3-Driven Pedagogical Agents to Train Children's Curious Question-Asking Skills. International Journal of Artificial Intelligence in Education (jun 2023). https://doi.org/10.1007/s40593- 023-00340-7

[2] Hojjat Abdollahi, Mohammad H. Mahoor, Rohola Zandie, Jarid Siewierski, and Sara H. Qualls. 2023. Artificial Emotional Intelligence in Socially Assistive Robots for Older Adults: A Pilot Study. IEEE Transactions on Affective Computing 14, 3 (July 2023), 2020–2032. https://doi.org/10.1109/taffc.2022.3143803

[3] Utku Günay Acer, Marc van den Broeck, Chulhong Min, Mallesham Dasari, and Fahim Kawsar. 2022. The City as a Personal Assistant. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 6, 2 (jul 2022), 1–31. https://doi.org/10.1145/3534573

[4] Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. 2020. Text-based emotion detection: Advances, challenges, and opportunities. Engineering Reports 2, 7 (may 2020). https://doi.org/10.1002/eng2.12189

[5] Israel Edem Agbehadji and Abosede Ijabadeniyi. 2020. Approach to Sentiment Analysis and Business Communication on Social Media. In Bio-inspired Algorithms for Data Streaming and Visualization, Big Data Management, and Fog Computing. Springer Singapore, 169–193. https://doi.org/10.1007/978-981-15- 6695-0_9

[6] Naveed Ahmed, Zaher Al Aghbari, and Shini Girija. 2023. A systematic survey on multimodal emotion recognition using learning algorithms. Intelligent Systems with Applications 17 (feb 2023), 200171. https://doi.org/10.1016/j.iswa.2022.200171

[7] Saima Aman and Stan Szpakowicz. [n. d.]. Identifying Expressions of Emotion in Text. In Text, Speech and Dialogue. Springer Berlin Heidelberg, 196–205. https://doi.org/10.1007/978-3-540-74628-7_27

[8] Kiavash Bahreini, Rob Nadolski, and Wim Westera. 2014. Towards multimodal emotion recognition in e-learning environments. Interactive Learning Environments 24, 3 (May 2014), 590–605. https://doi.org/10.1080/10494820.2014.908927

[9] David Baidoo-Anu and Leticia Owusu Ansah. 2023. Education in the Era of Generative Artificial Intelligence (AI): Understanding the Potential Benefits of ChatGPT in Promoting Teaching and Learning. SSRN Electronic Journal (2023), 22 pages. https://doi.org/10.2139/ssrn.4337484

[10] Nikola Banovic and John Krumm. 2018. Warming Up to Cold Start Personalization. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 1, 4 (jan 2018), 1–13. https://doi.org/10.1145/3161175

[11] Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The selfassessment manikin and the semantic differential. Journal of Behavior Therapy and Experimental Psychiatry 25, 1 (mar 1994), 49–59. https://doi.org/10.1016/0005- 7916(94)90063-9

[12] Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the Impact of Annotation Perspective and Representation Format on Dimensional Emotion Analysis. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. Association for Computational Linguistics, Valencia, Spain, 578–585. https://aclanthology.org/E17-2092

10.48047/jocaaa.2025.34.04.57

[13] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. Language Resources and Evaluation 42, 4 (nov 2008), 335–359. https://doi.org/10.1007/ s10579-008-9076-6

[14] Heng-Jui Chang, Shu wen Yang, and Hung yi Lee. 2022. Distilhubert: Speech Representation Learning by Layer-Wise Distillation of Hidden-Unit Bert. In ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. https://doi.org/10.1109/icassp43922.2022.9747490

[15] Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. SemEval-2019 Task 3: EmoContext Contextual Emotion Detection in Text. In Proceedings of the 13th International Workshop on Semantic Evaluation. Association for Computational Linguistics. https://doi.org/10.18653/v1/s19-2005

[16] Vishal Chudasama, Purbayan Kar, Ashish Gudmalwar, Nirmesh Shah, Pankaj Wasnik, and Naoyuki Onoe. 2022. M2FNet: Multi-modal Fusion Network for Emotion Recognition in Conversation. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE. https://doi.org/10.1109/cvprw56347.2022.00511

[17] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. 2019. What Makes a Good Conversation?. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM. https://doi.org/10.1145/3290605.3300705

[18] Bertrand David, Rene Chalon, Bingxue Zhang, and Chuantao Yin. 2019. Design of a Collaborative Learning Environment integrating Emotions and Virtual Assistants (Chatbots). In 2019 IEEE 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD). IEEE. https://doi.org/10.1109/ cscwd.2019.8791893

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.48550/ARXIV.1810.04805

[20] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-Powered Conversational Agents. In International Conference on Learning Representations. https://openreview.net/ forum?id=r1l73iRqKm