ISSN 2278-3091
Volume 13, No.3, May - June 2024
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCS1E/static/pdf/file/ijatcse071332024.pdf
https://doi.org/10.30534/ijatcse/2024/071332024

Empathy in AI: Developing a Sentiment-Sensitive Chatbot through Advanced Natural Language Processing

Dr.Mabrouka Abuhmida[1], Md Johirul Islam[2], Dr. Wendy Booth[3]
[1]University of south Wales filiation, UK, Mabrouka.abuhmida@southwales.ac.uk
[2]Coventry University, UK, mdjohirulislam545@gmail.com
[3]University of south Wales, UK, wendy.booth2@southwales.ac.uk

Received Date : April 17, 2024 Accepted Date: May 19, 2024 Published Date: June 06, 2024

ABSTRACT

This paper explores the development of an empathetic chatbot that utilises advanced natural language processing (NLP) techniques for accurate emotion detection and response generation. This paper also explores the impact of the use of a hybrid model that recognises and categorises emotional states from textual input. Using generation models like DialoGPT, fine-tuned with empathetic dialogue datasets. The proposed chatbot's architect performance was evaluated using various metrics, and the results show advantage over existing models.

Key words : Emotional Intelligence, Response Generation, Emotion Recognition, Empathetic Chatbot.

1.INTRODUCTION

The interaction between humans and artificial intelligence (AI) through conversational agents has rapidly evolved, emphasising the need for more sophisticated and empathetic communication. Recent advancements in NLP and deep learning have enabled the development of chatbots that can understand and appropriately respond to human emotions. This paper details the creation of an emotional sentiment chat-bot designed to detect and generate empathetic responses. By integrating a variety of machine and deep learning algorithms for emotion recognition and employing cut-ting-edge NLP models for response generation, this chatbot aims to bridge the gap between human emotional needs and AI capabilities. The methodology includes using multiple empathetic dialogue datasets and various NLP frameworks, offering a comprehensive approach to empathetic chatbot development.

2.RELATED WORKS

Sentiment analysis and chatbots that are empathetic have a lot of potential in mental health support, where it's very important to accurately identify and respond to users' emotions. Existing research in this field has highlighted the significance of technology in enhancing traditional mental health interventions and offering accessible help to persons in need. An in-depth review of research findings between emotional support and physical health was conducted by [1], and this study investigates the connection between social/emotional support and health out-comes based on mortality and morbidity. Several studies have been undertaken to evaluate the effectiveness of chatbots in giving emotional support. The authors of the research [2] completed an experiment regarding the performance of chatbots compared to human partners, and the result of this experiment provides a clear idea that the outcome remains consistent for both chatbots and humans. There is no difference, whether it is a chatbot or a human, to create emotional, psychological, and relational effects. According to [3], emotional support from the chatbot and human partners can reduce stress and worry, and this study provides several literature reviews regarding the importance of emotional support that gives valuable insight into emotional support and its impact on humans during stressful times. Aharoni and Fridlund (2007) [4] examined how people react when the interviewer is either a computer or a human during the interview and found that participants' thoughts and feelings were the same no matter who took the interview.

Chatbots provide a viable medium for mental support because of their accessibility and capacity to engage users in communication using natural language. Numerous researchers have studied chatbots and released significant findings regarding their evolution from early development to the sophisticated chatbots we use today. A brief overview of different chatbot types, the application of chatbots, and the definition of chatbots with in-depth concepts was conducted by [5]. A comprehensive analysis of early development chatbots has been completed since 1950 when Alan Turing proposed the Turing test to modern chatbots [6]. Different types of chatbot have been discussed base, such as knowledge domain, human aid, build methods, and generation methods. There are two types of chatbots: rule-based and machine learning-based chatbots [7]. In the research presented in [8] the authors give a detailed overview of existing chatbots that are used in educational fields, and the result shows that the

chatbots being used now have the potential impact on students' learning in different ways.

Chatbots for mental health can aid in providing therapy and training, and one of these is called ELIZA, which was developed by [9]. It is considered one of the most well-known and influential chatbots and can conduct human-like conversations. Eliza works on simple keyword matching/pattern matching and trans-formation rules. Another mental health support chatbot called SERMO uses cognitive behaviour therapy (CBT); this chatbot supports people who have mental illnesses in regulating their emotions, feelings, and thoughts [10].

Most early-development chatbots are rule-based, and rule-based chatbots interact with users based on predefined rules and decision trees. Their ability to manage com-plex queries is limited, but they are helpful for simple tasks and frequently asked questions. Researchers focused on advanced NLP techniques to address the limitations of early development chatbots. In research [11], the authors developed an empathetic chatbot called CAiRE following the process of transfer learning, which was first introduced in the study [12] that fine-tuned a pre-trained model for various purposes such as dialogue language modelling, emotion identifying from dialogue and response prediction. Another empathetic chatbot called Virtual Hope was proposed by [13]. The pre-trained model DialoGPT was utilised and fine-tuned on both the Empathetic Dialogue (ED) dataset and a custom well-being dataset in this study.

Traditional systems were programmed to recognise specific patterns and respond accordingly, and they failed to understand the significance of the context. Neural network-based architecture for a chatbot is more potential to make chatbots for generating responses compared to the rule-based technique [14]. An advantage of neural networks is their ability to learn patterns and relationships in vast amounts of data. They can evaluate context, interpret sentiment, and respond to user emotions and requirements.

Transformer architecture has made significant changes in the field of NLP. The Google researchers [15] introduced this novel architecture for ma-chine translation, and this architecture used an attention mechanism instead of a Recurrent Neural Network (RNN) and convolutional networks (CN) with encoder-decoder structures. The Microsoft researchers [16] introduced this DialoGPT model, which was fine-tuned on 147M Reddit conversations from 2005 to 2017. This DialoGPT model can create a response like a human, and the response of this model is more relevant, informative, and context consistent. The Hugginface researchers [12] introduced an innovative approach called Transfer-Transfo for a dialogue system that integrates transfer learning with a transformer model. The fine-tuning process involves utilising a multi-task objective that integrates unsupervised prediction tasks, and the fine-tuned model performs better than other conversational models, such as informational retrieval and seq2seq models. Rashkin et al. (2018) [17] presented empathic dialogue datasets and performed a comparative examination of retrieval architecture and generative architecture.

Their findings showed that the generative architecture outperformed alternative methods. The research [18] applied transfer learning using Di-aloGPT to generate Swedish dialogue: different open-source conversational datasets and one English dataset used for fine-tuned DialoGPT. The Facebook researchers [19] developed a large-scale model that can respond to user queries more engagingly, and the generated response is more reliable and human-like; it is also called Blenderbot. It is an opendomain chatbot that can respond accurately to user queries, and the whole training process is done using ParlAI frameworks. In [20], google researchers present an advanced multi-turn chatbot (open domain) called Meena that was trained using social media public domain con-versation that contains 2.6B parameters, and introduced human evaluation metrics, namely sensibleness and specificity average. Meena uses the Evolved Transformer (EV) seq2seq model, a type of Transformer architecture.

3.RESEARCH METHODOLOGY

The method used to make the emotional sentiment chatbot comprises two main parts: detecting emotions and generating empathetic responses. Seven algorithms were chosen and implemented in practice to tackle the emotion recognition task. These algorithms were selected based on their appropriateness for the task and encompassed a range of techniques, including machine learning and deep learning architectures. A hybrid emotion detection model has been developed by integrating the results of the distinct algorithms. At the same time, a cutting-edge method that has proven successful in natural language processing tasks—transformer architecture—was used to generate empathetic responses. The empathic response-generating model was trained using three empathetic dialogue datasets.

3.1 Data Collection

There are two main types of data available for public use: domain-specific and open-domain data. For this task, we used domain-specific data. This work searched various platforms for my desired datasets in research articles, Kaggle, GitHub, and online resources. Total of seven datasets are used that are publicly available for research and are available on the Hugginface (transformer library) website, and the rest of them are from the GitHub link that the researchers provide in research papers. Among seven datasets four datasets selected for emotion detection: ISEAR, EDOS, Emotion, and Multiclass emotional model datasets. For the empathetic response generation task, this work used three datasets: EmpatheticDialogue (ED), ProsocialDialogue (PD), and DailyDialogue (DD). EmpatheticDialogue (ED) dataset was proposed by Rashkin et al. (2018) [17], the Facebook AI researchers; it contains around 25 thousand conversations

grounded with emotional situations collected from 810 participants.

Kim et al. (2022) [21] first introduced this ProsocialDialogue dataset, and it is the first multi-turn large dataset used to respond to content that is problematic following social standards. It has 58 thousand conversations, 331 thousand utterances, RoTs 160 thousand and 497 thousand reasons, and safety labels. This dataset is publicly available for research on the Hugginface (transformer library) website. In [22], the researchers created DailyDialogue multi-turn high-quality dataset, which covers a wide range of related topics in our daily conversation. This dataset is made by raw data collected from the website where people practice English dialogue for everyday communications. EDOS dataset developed by [23] is one of the large datasets containing 1 million dialogues with 32 emotion that is fine-grained, with 8 intents for empathetic response and neutral. ISEAR dataset was introduced by [24], and the data was collected from a group of psychologists worldwide for the ISEAR project. The emotion dataset was introduced by [25]; it is a Twitter dataset containing six fundamental emotions: sadness, love, surprise, fear, joy and anger and it was developed by [26]. This dataset combines three datasets, namely DailyDialog, ISEAR and Emotion-stimulus. Table 1 provides a quick overview of the datasets used in this project.

Table 1: Dataset overview

Dataset Name	Authors	Year	Emo detection	Response Gene
Empathetic Dialogue	Rashkin et al.	2018	Yes	Yes
Prosocial Dialogue	Kim et al.	2022	No	Yes
Daily Dialogue	Li et al.	2017	No	Yes
ISEAR	Scherer and Wallbott	1994	Yes	No
EDOS	Welivita, Xie and Pu	2021	Yes	No
Emotion dataset	Saravia et al.	2018	Yes	No
Multiclass Emotional Model Dataset	Noramiza, Aznida and Aziah	2021	Yes	No

3.2 Data Preprocessing

Data preprocessing is one of the crucial parts of building any machine learning. It is an essential part of the data cleaning process because the quality of data has a substantial impact on various approaches, and it involves Tokenisation, Removing Stop-words, Stemming, Lemmatisation, Pos-Tagging, and normalisation [27]. As this project combines different datasets to detect emotion and generate emotional responses, this work conducts some preprocessing steps to make the proposed model more reliable and faster. After combining the datasets, the total unique emotion labels were 47 categories. This work used categories of emotional labels such as: joyful, sad, hopeful, afraid, caring, angry, proud, guilty, terrified, faithful, impressed, lonely, annoyed, wishing, and surprised.

The sequence of sentence length was fixed to a max sequence length was set to 50 words. All duplicate values were removed from the dataset. The combined dataset was imbalanced, and the distribution of emotional labels was not fixed. For example, the value count for the inspirational label wishing was 67067, while the value count for the emotional label sad was 22413. We applied data balancing techniques such as Resampling. Finally, tokenisation was used for the bi-directional LSTM model. Figure. 1 shows the text preprocessing stages that this work follows.

Figure. 1: Data Preprocessing Steps

3.3 Emotion Recognition Algorithm Design

A variety of machine learning (ML) and deep learning algorithms were employed for emotion detection. Five machine learning and two deep learning techniques have been applied for this task; a hybrid model has been built to increase the chatbot's accuracy for correct emotion recognition by leveraging all emotion detection algorithms. The process of converting raw text into a format that machine learning algorithms understand is called feature extraction, which generates a feature matrix. In the feature matrix, each raw represents a sentence , and the words are represented in columns. There are many methods used for feature extraction, such as Bag of words, Ngram, TFIDF, and word embedding [27]. This work used the term Frequency-inverse document frequency (TF-IDF) method for sentiment classification TF-IDF reflects the significance of a word in a document in relation to its frequency throughout the corpus, and the concept of TF-IDF is to provide more weight to the informative and distinctive words in the corpus. And the Glove word embedding is used for the content generation in the neural network. Glove captures global corpus statistics directly [28].

This study has utilised compared different methods for sentiment analysis and response generation. These methods include traditional machine learning algorithms such as SGDClassifier, MultinomialNB, RandomForestClassifier, LinearSVC, and KNeighborsClassifier. Additionally, deep learning architectures, namely Bi-directional LSTM and a pre-trained bart-large-mnli model.

Each algorithm serves a specific purpose in the classification process. For instance, Naive Bayes (MultinomialNB) is highlighted for its simplicity and widespread usage in text classification based on word distribution. LinearSVC is noted for its effectiveness in binary and multiclass problems, particularly well-suited for text classification due to its ability to find optimal hyperplanes. K-Nearest Neighbors (KNN) is acknowledged as a basic yet effective algorithm for text classification, relying on nearest neighbors and distance

metrics [29]. RandomForestClassifier, an ensemble learning method using decision trees, is also employed for its effectiveness in text classification tasks.

In the realm of deep neural network, the Bi-directional LSTM [30] is utilised to capture sequential data and solve vanishing gradient problems commonly encountered in recurrent neural networks. Additionally, a pre-trained bart-large-mnli model is employed, indicating the use of transfer learning to leverage pre-existing knowledge for the task at hand [31].

These methods are selected and combined to create a hybrid model for emotion detection, with the most accurate machine learning model chosen from a selection of five others. The overall process for emotion recognition is depicted in Figure. **2**, indicating the comprehensive approach taken in the study.

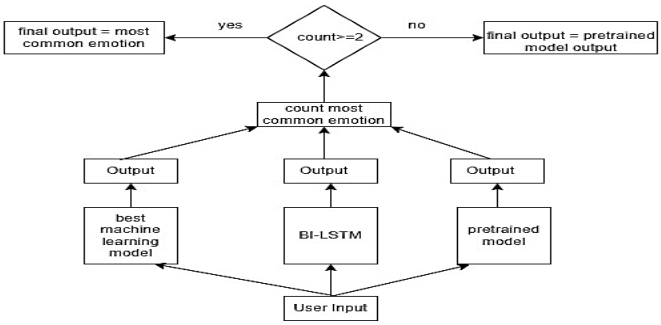

Figure. 2: Proposed System Architecture

3.4 Response Generation Model Selection

The choice of response generation model is crucial for effective chatbots. Quality and acceptance rely on model selection. We opted for the original transformer model and DialoGPT. The transformer architect is a pivotal advancement in natural language processing, capable of handling sequential data through multi-head attention mechanisms. Proposed by Google in 2017, it comprises encoder, decoder, feedforward, positional encodings, and multi-head attention.

DialoGPT, is an advanced language model developed by Microsoft researchers, as the basis for generating responses. DialoGPT was fine-tuned utilising empathetic data to improve its ability to generate empathic responses. The architecture for generating empathetic responses is designed to generate contextually meaningful and emotionally intelligent responses in natural language.

3.5 Generative chatbot and emotion classifier

An emotion classifier uses the user input and detects the user's emotional state to identify emotion. At the same time, an empathetic chatbot uses the same user input and generates a sympathetic response. This proposed architecture for developing an empathetic chatbot helps bridge the gap between AI and human emotions, leading to empathetic and meaningful interactions is highlighted in

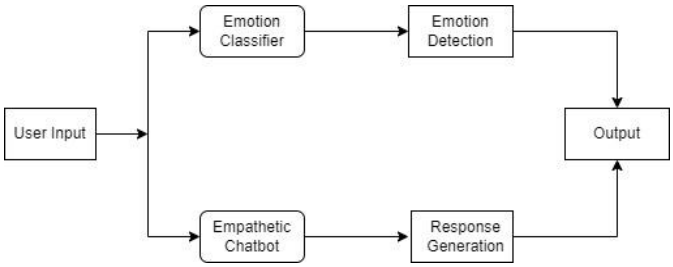

Figure 3: Proposed System Architecture

3.6 Training and Hyperparameter Tuning

This work contains three main components and undertaken an in-depth implementation of all these components using various popular Deep learning and Machine learning libraries and frameworks, namely TensorFlow, Karas, Pytorch, Transformers, SKlearn, NumPy, Pandas.

The dataset was split into train and test, where “80%” of the data is used for train and “20%” for test. To implement Bidirectional LSTM, glove word embedding is used as a feature extraction method. Tokenisation, Padding, and Text to Sequence are all methods used to prepare data for training. This study defined input shape = 60, the number of emotions, or output = 15 to implement the bidirectional LSTM. The embedding layer is used for feature extraction, and the SpatialDropout layer is utilised for regularisation, where 0.2 is the dropout rate. The Conv1D layer is applied to capture the input data pattern with 64 filters, filter size “5”, and activation function “relu”. In the Bidirectional layer “64” LSTM units are utilised with a dropout of 0.2. Two hidden dense layers, one dropout layer and one output dense layer, are used, where “512” is the hidden units. The “softmax” activation function was applied for the output dense layer, and for training, the total epoch was set to “30” with batch size “256”.

The transformer is considered one of the most important inventions of deep learning and natural language processing. Google researchers introduced this architecture, and the primary intention of this architecture was machine translation. The secret success of this architecture is the attention mechanism that allows the model to find the relationship between each word. The key components of this architecture are multi-head attention, positional encoding, encoder layers, encoder, decode layers, decoder, and feedforward networks. A diverse set of hyperparameters is used to implement this work, and these hyperparameters play an essential role in the training process and boost the model's performance. As we aimed to implement the original transformer architecture, several hyperparameters were used from the original paper to reduce the computational cost and make the model faster; some parameters have been changed compared to the original paper. This implementation used the hyperparameter, namely the number of encoder and decoder layer = 2, model dimension = 512, number of hidden units = 128, batch size = 64, number of heads = 8, and maximum length = 60, epoch = 40, and learning

rate = 1e-4 where in the original paper they used number of hidden units = 2048, encoder and decoder layer = 6, number of heads = 8, model dimension = 512.

4 RESULTS AND ANALYSIS

The results section thoroughly summarises the findings derived from assessing the emotion detection and empathetic response generation model of the emotional sentiment chatbot. The analysis of these results demonstrates the model's efficacy in effectively detecting and categorising emotions in user input, showcasing its capacity for identifying nuanced emotional states. It employs various techniques and metrics to evaluate the model's performance on a particular task.

4.1 Evaluation Metrics

Selecting an evaluation metric relies on the type of problem, and there are several types of evaluation metrics, namely, classification metrics, regression metrics, clustering metrics, and natural language processing metrics. Accuracy, precision, recall, f1-score, and confusion metrics are all used as classification metrics, and BLEU and perplexity are used as NLP metrics. This work unitises several classification metrics to evaluate the performance of emotion detection (summarised in Table 2) and uses perplexity as an evaluation metric for response generation.
Perplexity evaluates the quality of a language model by measuring how well the model predicts a sample of text. [20]. The lower perplexity score is considered best for predicting the next words [18]. The following table summarises the other evaluation metrics used.

Table 2: Evaluation Metrics Summary

Metric	Description
Confusion Matrix	Evaluation tool in binary or multiclass classification tasks; provides insight into how model predictions align with actual values.
Accuracy	Percentage of accurately predicted instances out of the total instances in a dataset; reflects overall performance of a model's predictions.
Precision	Measures true positives compared to all positive predictions; indicates the model's ability to correctly identify positive cases.
Recall	Measures true positives among all actual positives; represents the model's ability to capture all relevant instances.
F1-Score	Harmonic mean of precision and recall; combines both precision and recall, providing a balanced measure of a model's performance, particularly useful when there is an imbalance.

4.2 Performance Analysis of The Emotion Recognition Algorithm

This work comprehensively analysed emotion detection algorithms using various machine learning and deep learning approaches. These wide ranges of algorithms aim to recognise and categorise emotions from textual data automatically. Table 3 provides an overview of evaluation scores for emotion detection algorithms.

Table 3: Overview of Evaluation Metrics for 15 Categories

Model	Accuracy	Precision	Recall	F1-score
Multinomial NB	0.70	0.72	0.70	0.70
SGD	0.70	0.71	0.70	0.70
KNeighbors	0.61	0.63	0.61	0.61
Random Forest	0.66	0.67	0.66	0.66
Linear SVC	0.75	0.75	0.75	0.75
Bidirectional LSTM	0.77	0.78	0.77	0.77

Based on the analysis of these six models, the Bidirectional Long Short-Term Memory (BiLSTM) and Linear Support Vector Classifier (Linear SVC) models were found to have more outstanding accuracy performance than other models. The score of all evaluation metrics in Table 3 is for 15 different emotion categories. In order to test the differences in these scores, the emotion categories were reduced from 15 to 8 for investigation purposes. After reducing the emotion categories, all models provide higher accuracy. Figure 4 gives an overview of the evaluation score for all models after reducing the emotion categories from 15 to 8.

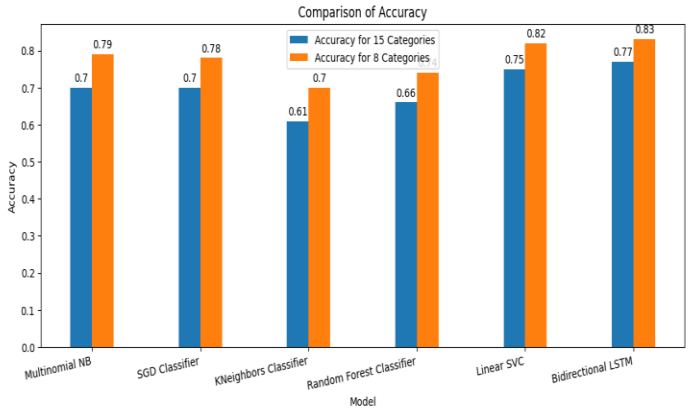

Figure 4: Comparison of Accuracy

4.3 The Response Generation Model And Integration

This work applied a Transformer architecture and fine-tuned the DialogPT model to boost its response generation capabilities. Perplexity was used as a metric to evaluate the effectiveness of the models. After applying perplexity metrics

to the model, we found that the model that used transformer architecture gives lower perplexity compared to DialoGPT, which means the response of the transformer model is more contextual and coherent. Table 4 provides an overview of the perplexity score.

Table 4: Perplexity score

Model	Perplexity
Transformer Model	1.78
DialoGPT Small	5.29
DialoGPT Medium	4.92

Among the three models, the lowest perplexity is 1.78 for the transformer model, which indicates that this model is more certain to predict the next word more accurately than DialoGPT small and medium. Table 5 shows a sample response of these model for the same text.

Table 5: Sample Response Generation

Text	DialoGPT-Small	DialoGPT-Medium	Transformer
Hi how are you?	I m fine, thanks. how are you?	I'm fine. How about you? Are you busy?	I'm fine. How about you? What's going on?
can I talk to you for a moment?	sure. what do you need?	I'd love to, but I'm afraid I can't.	sure. What's up?
I've been second-guessing myself and feeling like I'm not good enough.	I hope you can work out for and look after yourself.	I understand. I hope you find a way to improve.	You should be proud of yourself for doing well.
One of my close friends leave me without any reason	I am sorry for that. Why do you think they left you?	Why do you think that they don't want to go with you?	I am sorry to hear that. Why do you think they left you without any good reason?

After completing the response generation task, this work integrates the emotion detection hybrid model with the response generation model (Transformer), where the hybrid model ensures the correct emotion from user text, and the response generation model provides an appropriate contextual and coherent response based on user emotions.

4.4 Comparison with existing Emotional AI Chatbots

This work compares the performance of the proposed chatbot architect to other similar chatbots. The proposed chatbot architect perplexity scores were lower than other models. To compare the performance, we used three existing advanced AI chatbots, CAiRE, VHope, and another [17]. The lowest perplexity scores of CAiRE and VHope chatbots were 13.32 and 9.64, respectively, whereas in [17] the lowest perplexity was 21.24 among the four models they built. Table 6 shows the Perplexity score of all models.

The lowest perplexity scores show that this emotional sentiment chatbot is better than other models at generating appropriate responses.

Table 6: Comparison with existing Emotional AI chatbots

Model	Perplexity Score
Transformer (This work)	*1.78*
CAiRE (Lin et al., 2020)	*13.32*
VHope (Beredo and Ong, 2022)	*9.64*
Fine-tuned BERT (Rashkin et al.)	21.24

5 CONCLUSION

The emotional sentiment chatbot developed in this study demonstrates a significant stride towards creating AI that can interact humanely with users. The hybrid model combining machine learning techniques and transformer architecture for emotion recognition and response generation has proven highly effective. The chatbot's performance, measured by metrics like accuracy and perplexity, outperforms conventional models and showcases its capability to engage in meaningful and supportive dialogues. These findings reinforce the viability of empathetic chatbots in practical applications, such as mental health support, and suggest directions for future research in making AI interactions more relatable and supportive. The success of this chatbot underscores the potential of integrating complex NLP techniques to enhance the emotional intelligence of AI systems, paving the way for more nuanced and genuinely empathetic technological interactions.

In conclusion, future research directions are proposed to enhance the development and effectiveness of empathetic chatbots. These include refining emotion recognition algorithms using advanced deep learning techniques, expanding the diversity and size of training datasets, and testing the chatbots in real-world applications like customer service and mental health support. Additionally, incorporating multimodal inputs and personalisation features could improve the chatbot's responsiveness and user engagement. Addressing ethical concerns, such as data privacy and bias mitigation, is also critical.

REFERENCES

[1] M. Reblin and B. N. Uchino, ‘Social and emotional support and its implication for health’, *Curr. Opin. Psychiatry*, vol. 21, no. 2, pp. 201–205, 2008, doi: 10.1097/yco.0b013e3282f3ad89.

[2] A. Ho, J. Hancock, and A. S. Miner, 'Psychological, Relational, and Emotional Effects of Self-Disclosure After Conversations With a Chatbot', *J. Commun.*, vol. 68, no. 4, pp. 712–733, 2018, doi: https://doi.org/10.1093/joc/jqy026.

[3] M.-H. Guo *et al.*, 'Attention mechanisms in computer vision: A survey', *Comput. Vis. Media*, vol. 8, no. 3, pp. 331–368, 2022.

[4] E. Aharoni and A. J. Fridlund, 'Social reactions toward people vs. computers: How mere lables shape interactions', *Comput. Hum. Behav.*, vol. 23, no. 5, pp. 2175–2189, 2007, doi: https://doi.org/10.1016/j.chb.2006.02.019.

[5] A. Gupta, D. Hathwar, and A. Vijayakumar, 'Introduction to AI chatbots', in *International Journal of Engineering Research and Technology*, 2020, pp. 255–258.

[6] E. Adamopoulou and L. Moussiades, 'An Overview of Chatbot Technology', in *IFIP Advances in Information and Communication Technology*, 2020, pp. 373–383. doi: https://doi.org/10.1007/978-3-030-49186-4_31.

[7] S. Meshram, N. Naik, M. VR, T. More, and S. Kharche, 'Conversational AI: Chatbots', in *2021 International Conference on Intelligent Technologies (CONIT)*, 2021. doi: 10.1109/conit51480.2021.9498508.

[8] J. Q. Pérez, T. Daradoumis, and J. M. M. Puig, 'Rediscovering the use of chatbots in education: A systematic literature review', *Comput. Appl. Eng. Educ.*, vol. 28, no. 6, 2020, doi: 10.1002/cae.22326.

[9] J. Weizenbaum, 'ELIZA—a computer program for the study of natural language communication between man and machine', *Commun. ACM*, vol. 9, no. 1, pp. 36–45, 1966, doi: 10.1145/365153.365168.

[10] K. Denecke, S. Vaaheesan, and A. Arulnathan, 'A Mental Health Chatbot for Regulating Emotions (SERMO) - Concept and Usability Test', *IEEE Trans. Emerg. Top. Comput.*, vol. 9, no. 3, pp. 1–1, 2020, doi: https://doi.org/10.1109/tetc.2020.2974478.

[11] Z. Lin *et al.*, 'CAiRE: An End-to-End Empathetic Chatbot', in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 13622–13623. doi: 10.1609/aaai.v34i09.7098.

[12] T. Wolf, V. Sanh, J. Chaumond, and C. Delangue, 'TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents', *ArXiv Prepr. ArXiv190108149*, 2019, doi: 10.48550/arxiv.1901.08149.

[13] J. L. Beredo and E. C. Ong, 'A Hybrid Response Generation Model for an Empathetic Conversational Agent', 2022. doi: https://doi.org/10.1109/IALP57159.2022.9961311.

[14] L. Wang *et al.*, 'CASS: Towards Building a Social-Support Chatbot for Online Health Community', *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW1, pp. 1–31, 2021, doi: 10.1145/3449083.

[15] A. Vaswani *et al.*, 'Attention is all you need', in *Advances in Neural Information Processing Systems*, 2017.

[16] M. Zhang, M. Li, J. Zhang, L. Liu, and H. Li, 'Onset detection of ultrasonic signals for the testing of concrete foundation piles by coupled continuous wavelet transform and machine learning algorithms', *Adv. Eng. Inform.*, vol. 43, p. 101034, Jan. 2020, doi: 10.1016/j.aei.2020.101034.

[17] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, 'Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset', 2018, doi: 10.48550/arxiv.1811.00207.

[18] T. Adewumi *et al.*, 'Småprat: DialoGPT for Natural Language Generation of Swedish Dialogue by Transfer Learning', 2021.

[19] S. Roller *et al.*, 'Recipes for building an open-domain chatbot', *ArXiv Prepr. ArXiv200413637*, 2020, doi: 10.48550/arxiv.2004.13637.

[20] D. Adiwardana *et al.*, 'Towards a Human-like Open-Domain Chatbot', 2020.

[21] H. Kim *et al.*, 'ProsocialDialog: A Prosocial Backbone for Conversational Agents', *ArXiv Prepr. ArXiv220512688*, 2022, doi: 10.48550/arxiv.2205.12688.

[22] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, 'DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset', in *International Joint Conference on Natural Language Processing*, 2017, pp. 986–995.

[23] A. Welivita, Y. Xie, and P. Pu, 'A Large-Scale Dataset for Empathetic Response Generation', in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021. doi: 10.18653/v1/2021.emnlp-main.96.

[24] K. R. Scherer and H. G. Wallbott, 'Evidence for universality and cultural variation of differential emotion response patterning: Correction', *J. Pers. Soc. Psychol.*, vol. 67, no. 1, pp. 55–55, 1994, doi: 10.1037/0022-3514.67.1.55.

[25] E. Saravia, H.-C. T. Liu, Y.-H. Huang, J. Wu, and Y.-S. Chen, 'CARER: Contextualized Affect Representations for Emotion Recognition', in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 3687–3697. doi: 10.18653/v1/D18-1404.

[26] H. Noramiza, A. Aznida, and A. Aziah, 'Multiclass Emotion Model Dataset'. 2021. doi: 10.5281/zenodo.5040202.

[27] P. Nandwani and R. Verma, 'A review on sentiment analysis and emotion detection from text', *Soc. Netw. Anal. Min.*, vol. 11, no. 1, 2021, doi: 10.1007/s13278-021-00776-6.

[28] J. Pennington, R. Socher, and C. Manning, 'Glove: Global Vectors for Word Representation', in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014. doi: 10.3115/v1/d14-1162.

[29] A. P. Jain and P. Dandannavar, 'Application of Machine Learning Techniques to Sentiment Analysis', 2017. doi: https://doi.org/10.1109/ICATCCT.2016.7912076.

[30] P. Baid, A. Gupta, and N. Chaplot, 'Sentiment Analysis of Movie Reviews using Machine Learning Techniques', *Int. J. Comput. Appl.*, vol. 179, no. 7, pp. 45–49, 2017, doi: 10.5120/ijca2017916005.

[31] P. Bahad, P. Saxena, and R. Kamal, 'Fake News Detection using Bi-directional LSTM-Recurrent Neural Network', *Procedia Comput. Sci.*, vol. 165, pp. 74–82, 2019, doi: 10.1016/j.procs.2020.01.072.