

# CARO: An Empathetic Health Conversational Chatbot for People with Major Depression

**Nidhin Harilal**

Undergraduate  
IIT Gandhinagar  
17110092

nidhin.harilal@iitgn.ac.in

**Rushil Shah**

Undergraduate  
IIT Gandhinagar  
17110135

shah.rushil@iitgn.ac.in

**Saumitra Sharma**

Undergraduate  
IIT Gandhinagar  
17110140

saumitra.sharma@iitgn.ac.in

**Vedanta Bhutani**

Undergraduate  
IIT Gandhinagar  
17110171

vedanta.bhutani@iitgn.ac.in

## Abstract

There has been a rise in the number of patients suffering from major depression over the past decade, proliferating mainly in teenagers, due to number of reasons. Although there are several counseling services available, most of the patients are reluctant and do not open up for it. The reluctance mainly originates from one's fear of disclosure of identity and the social stigma against mental-health problems. Conversational applications such as *chatbots* have been found efficient in overcoming alcohol addiction through continuous conversation and user monitoring.(Ren et al.) Effective treatments can tackle depression, but more than 90 percent of affected patients are not able to avail such treatments, mainly due to lack of resources and social stigma associated with mental-health problems. In this paper, we propose *CARO*, a chatbot application, which is capable of performing *empathetic conversations* and *medical advice* for people with major depression. *CARO* will be able to detect the context from the conversation, its intent and the associated emotions. Depending upon the resultant understanding developed from the features detected, *CARO* can generate an empathetic response or medical advice according to the intent of the user.

which is not an impressive figure.(of [Psychiatry, March 2017](#)) Mental well-being is an essential part of our life. Mental Diseases are generally hard to diagnose, and they are accompanied by symptoms like fatigue, insomnia, etc.([Psychiatry, May 2017](#)) In contemporary society, there is a stigma towards the problems associated with mental health, which is making it difficult for the people who are willing to share their problems from doing so.

Moreover, Less than 10% of patients have access to formal therapy for their illness.(et al., [Apr 2011](#)) Without proper consult and treatment, a person may fall deep into the depression well. Existing solutions suffer from monotonicity in the type of responses they provide, and they hardly provide any medical advice. An intelligent chatbot equipped with the knowledge of empathy as well as the medicine can serve as a customized, round the day therapy.

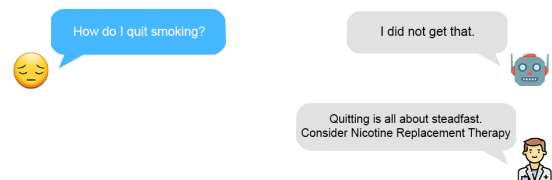


Figure 1: An example use of chatbot

## 1 Introduction

Cases of depression and other mental health problems have proliferated over the past decade due to various reasons. According to a report provided by the Government of India in 2018, the ratio of availability of psychiatrists to the people is 0.000077,

Our proposed solution, *CARO*, can serve as a constant companion for the patient by providing an empathetic response or medical advice. We see this as an opportunity to bridge the gap between

medical advice and general empathetic response, which, to the best of our knowledge, has not been addressed by anyone before. The task of generating an empathetic response or medical advice requires us to understand the intent and emotion of the given text input. To achieve this, we are training two different models on two separate datasets and later combining them with the help of intent.

In the future of chatbot based solutions, there is still much room for significant advancements.(Nuruzzaman and Hussain, 2018) Chatbots would be able to deal with most of the trivial and repetitive tasks better than ever. In the field of medical science, it can bring a radical change by providing a personalized user experience and providing health reports while monitoring the patient, and immediately notifying the doctor in the cases where utmost care must be taken. Moreover, in the near future, by means of breakthrough innovations, it will be possible to produce even the most complex results in real-time. It will help the incorporation of such technology not only in Smart-watches, Smart-TVs, etc. but also in day-to-day articles, without much increase in their cost.

## 2 Background

There has been a lot of work on conversational chatbots in last decades. None of attempts so far provides both medical advice and general empathetic response based on the context of conversation. Some existing studies shows that adding emotion information into dialogues improve user satisfaction. (Polzin and Waibel, 2000) Facebook AI generated a novel dataset(Rashkin et al., 2019b) consisting of conversations labelled with emotions. They also proposed a model for generating empathetic responses using a transformer based model.(Vaswani et al., 2017) They used a retrieval based architecture for generating responses. A retrieval based architecture involves two components, a generator and a retrieval model.(Swanson et al., 2019) The generator model generates multiple possible responses for a given context. Retrieval based model chooses the best response out of the generated response with given context. In order to account for emotion they prepended the emotion with sentences. Both generator and retrieval model are transformer based.

The work (Dongkeon Lee et al., 2017) discusses

a multi-modal approach to include emotion in the chatbots. They proposed to use various sources such as text, facial expressions, intonations, to determine the emotion of the input. Continuous observation of user can provide better information about the user's mental health.(Dongkeon Lee et al., 2017) Their work focusses on the task of combining various modes of determining the emotion. On the other hand, our work focuses on identifying emotion from the text and producing empathetic responses accordingly.

XiaoIce (Zhou et al., 2018) is uniquely designed AI companion with an emotional connection to satisfy the human need for communication, affection, and social belonging. XiaoIce take into account both intelligent quotient (IQ) and emotional quotient (EQ) in system design, cast human-machine social chat as decision-making over Markov Decision Processes (MDPs), and optimize XiaoIce for long-term user engagement, measured in expected Conversation-turns Per Session (CPS).(Zhou et al., 2018)

They (Serban et al., 2015) extended the hierarchical recurrent encoder-decoder neural network to the dialogue domain, and demonstrates that their model is competitive with the state-of-the-art neural language models and back-off n-gram models. They also investigated the limitations of their approach and similar approaches, and showed how its performance can be improved by bootstrapping the learning from a larger question-answer pair corpus and from pretrained word embeddings.(Serban et al., 2015)

The work (Li et al., 2019) developed a conversation content generation model that combines reinforcement learning with emotional editing constraints to generate more meaningful and customizable emotional replies. The model combines multi-task learning with multiple indicator rewards to comprehensively optimize the quality of replies.(Li et al., 2019) Experiments shows that their model significantly enhance the logical relevance and emotional relevance of the replies. (Li et al., 2019)

They (Pamungkas, 2019) did a comparison survey of all the chatbot works that in some way relate to empathetic response. The work (Cattania et al., 2019) built a modular framework to

facilitate and accelerate the realization and the maintenance of intelligent Conversational Agents with both rational and emotional capabilities rule-based approach with capability to simulate emotion.(Pamungkas, 2019)

Many medical chatbots follow a questionnaire-based approach, where they ask the user yes/no questions based on their symptoms. Products like GYANT (Gyant, 2016), is one such chatbot modeled on hand-crafted finite-state automata. The pre-defined rules dictate the conversation flow in such chatbots. Apart from open-source chatbots, many commercial products (e.g. GYANT) also use such approach. Our chatbot capitalizes on data of recorded medical conversation to answer the user queries.

Their work (Kevin A.) describes an AI health chatbot which works on "user inputs". Their digital therapy service calculates various metrics, such as scores and milestone determinations, to measure the customer's progress.

They (Hoermann et al., 2017) discussed the current evidence for the feasibility and effectiveness of online one-on-one mental health interventions that use text-based synchronous chat. (Divya et al., 2018) One such attempt to generate one to one chat is the work (Divya et al., 2018) describing a chatbot which uses a FSM logic to achieve an accurate diagnosis. The logic for state transitions were hard coded. Natural language generation templates were used to get responses from the user.

Our aim is not to provide a full solution but to guide the user in the right direction. We envision our chatbot to understand and help the user to achieve a healthy life.

### 3 Methodology

Our proposed model is an ensemble of two models, one is a medical advice generator, and the second is a general empathetic conversation generator. The given user-utterance gets directed into one of these models based on its specific defined intent. The reason behind our introduction of this intent feature is that it will guide the given

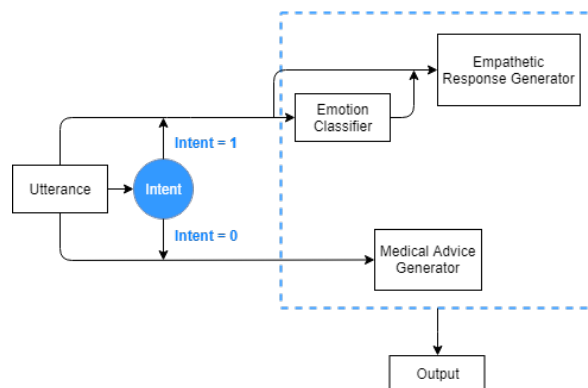


Figure 2: Proposed pipeline for response generation

user-utterance in identifying whether it is a part of a general conversation or a piece of medical question. Since the value of intent is a boolean (indicating its category as described above), therefore we built it as a binary classifier that was trained on data consisting of medical question answers and empathic dialogue responses.

Therefore, when given a text, first, its intent is classified as 1 or 0, based on which it goes into either of the defined models. Figure 2 describes our overall proposed pipeline. The first model, i.e., medical advice generator, follows an LSTM(Hochreiter and Schmidhuber, 1997) based architecture trained on a dataset containing different medical question answers. Whereas the other model, which generates empathetic responses, is trained on a dialogue dataset containing empathic conversations. The model also takes into account the previous two utterances along with the emotions associated with the current utterance. Emotions are incorporated by prepending them with the current utterances. The prepended emotions help in involving. And, introducing the previous two utterances along with the current one in generating the response helps in maintaining the context of the conversation. This model also follows a feedback mechanism where the final word is generated on word-by word basis at a given instance.

### 4 Datasets

We took a data-driven approach for both generation and classification tasks. We have used Facebook AI Empathetic Dialogue (Rashkin et al., 2019c) dataset and Medical Question Answering dataset (Nielsen, 2017).

## 4.1 Facebook AI Empathetic Dialogue Dataset

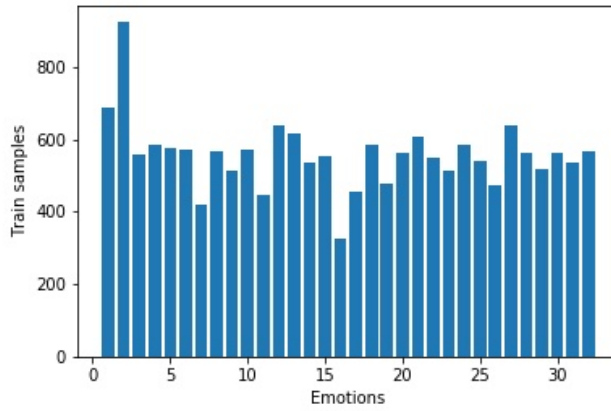


Figure 3: Distributions of Emotions in Facebook Empathetic Dialogue Dataset (Rashkin et al., 2019c)

Empathetic Dialogue(ED)(Rashkin et al., 2019a) dataset consists of 24,850 empathetic dialogue conversation. Each dialogue is grounded in a specific situation where a speaker was feeling a given emotion, with a listener responding. In total, there were 32 emotions, and on average, per class had less than 1000 number of examples. During the analysis of this data, We observed that there were many emotions which were very similar and could actually be grouped together, for example: "angry" and "furious" not necessarily means the same but could be treated as one. Based on this hypothesis, We grouped all the 32 emotions into eight groups possessing similar meanings. One other benefit of doing this was that we obtained an increased number of examples per class with a balanced distribution which guided in the training process.

## 4.2 Medical Q/A Dataset

Medical Question Answering dataset is a web-scraped ensemble of question-answers pairs consisting of Medical advices from various medical counselling forums such as *eHealth Forum*(eHealthForum, 2019), *HealthTap*(HealthTap, 2019), *WebMD*(WebMD, 2019) etc. During the analysis of this data, we observed that some of the questions in this dataset were either too long or were too specific and complex in terms of the

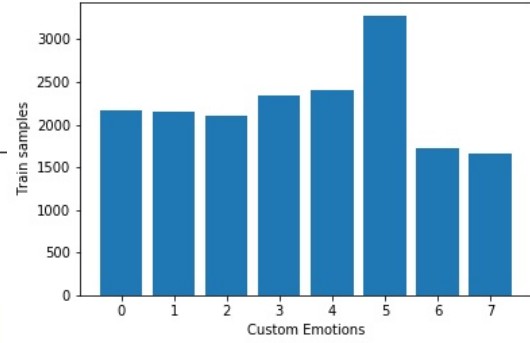


Figure 4: Distribution of proposed Emotion classes

medical conditions, it described. Since, most of the questions were tagged with their relevant labels, we utilized those labels to find and extract out the one that were related to medical symptoms related to the condition of depression. Some of the examples of such tags involved Insomnia, Weight Loss, Fatigue etc. The reason behind hand-picking questions with such tags were because If the affective overlap between both of the used dataset were not much than the test queries that the our model learns would be of completely different distributions. So, After all the question-answers related to that tag were filtered out, the dataset now consisted of 35,294 question-answer pairs.

## 5 Architecture

The complete architecture involved four different models for the four separate , Intent Classification, Emotion Classification, Empathetic Response Generation, and Medical Response Generation.

### 5.1 Baseline

Seq2Seq model proposed in (Sutskever et al., 2014), is a major achievement in translation-related tasks. We utilised this model to generate an empathetic response. The emotions extracted by the emotion classifier were appended at the beginning of the context sentence to generate an empathetic response. This model consists of the Encoder-Decoder model, where each Encoder and Decoder block consists of one or more LSTM/GRU units.(Chung et al., 2014) Encoder block captures the context features. The output of the encoder is fed as an input to the decoder block. We achieved a BLEU score of 0.06 when

Table 1: Proposed Emotion classes

Grouped Emotions	Different Emotions
Emotion 0	excited, surprised, joyful
Emotion 1	afraid, terrified, anxious, apprehensive
Emotion 2	disgusted, embarrassed, guilty, ashamed
Emotion 3	angry, annoyed, jealous, furious
Emotion 4	faithful, trusting, grateful, caring, hopeful
Emotion 5	sad ,disappointed, devastated, lonely, nostalgic, sentimental
Emotion 6	proud, impressed, content
Emotion 7	anticipating, prepared, confident

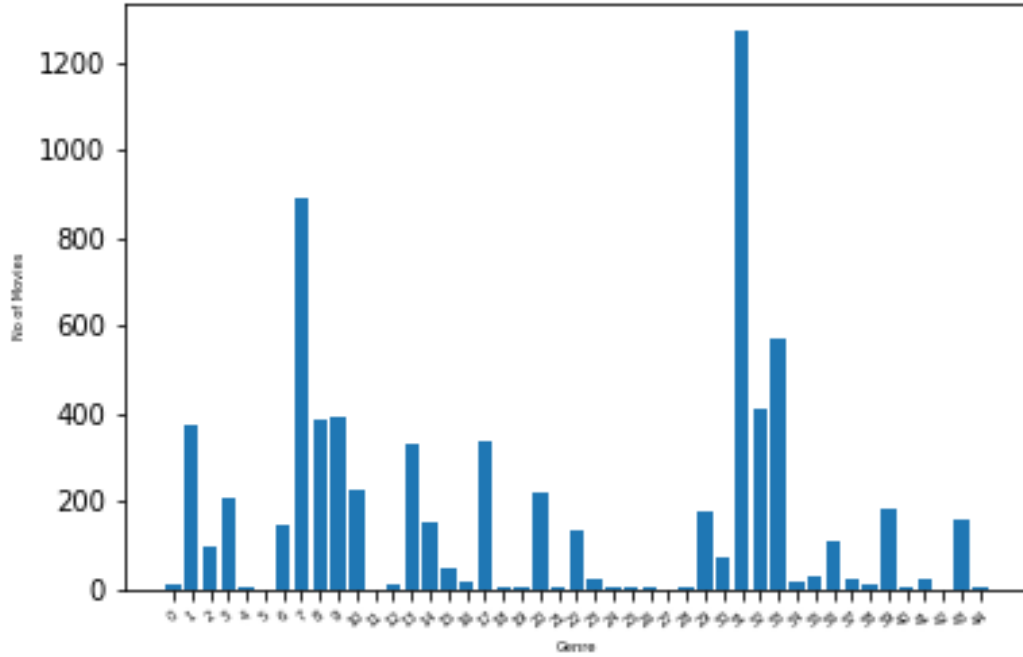


Figure 5: Distribution of number of Medical Questions corresponding to each category. Some of the prominent categories were psychitric, sleep hygeine, depression, mental-health, addiction, consulting, smoking, zoloft, schizophrenia, anxiety, suicidal, stress, pain, family, relationship

we trained this model on Facebook AI’s Empathetic Dialogues Dataset.

## 5.2 Classifiers

Long Short-Term Memory (LSTM) networks are used in our architectures as the core unit. It helps

in preserving both long and short term dependencies. As our input is a sequence of words that might have a dependency on each other, using LSTM helps the model to figure out the context of the utterance.

We used an Embedding Layer, which is initialized using Glove embeddings to convert the input

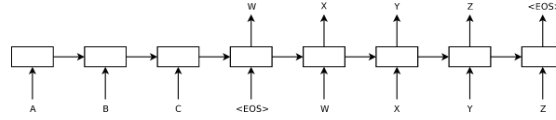


Figure 6: Baseline Model Seq2Seq source (Sutskever et al., 2014)

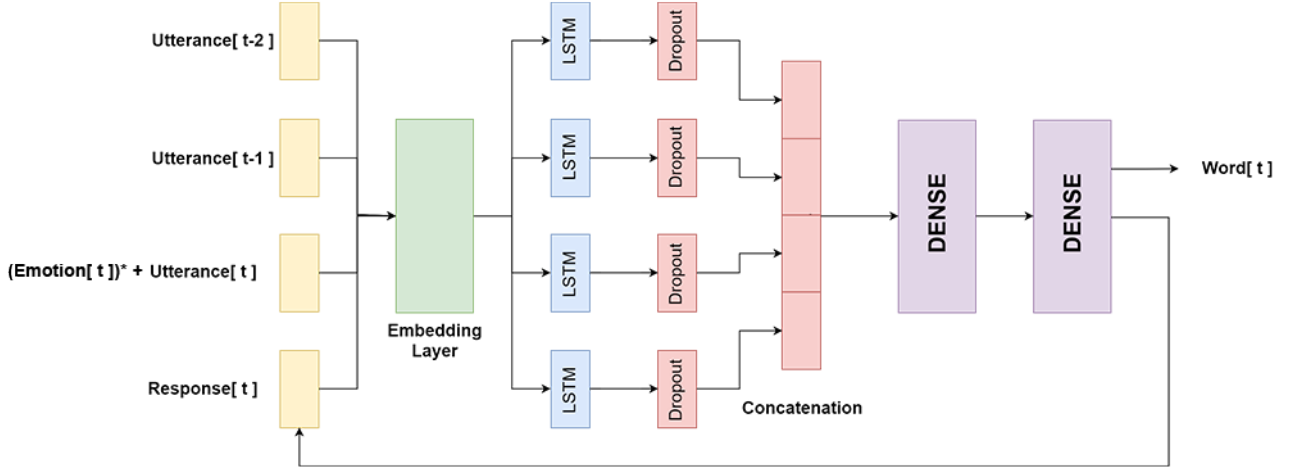


Figure 7: Parallel LSTM based Response Generation Architecture

sentence in the form of an embedding vector.

### 5.3 Intent Classifier

The Intent Classifier encompasses an LSTM Unit consisting of 300 cells, which generates a decoded sequence of the input embedding. This decoded sequence is passed on to a Dense Layer consisting of 100 units. As there are only two intents, medical advice, and empathetic response, the output is passed to a Dense layer with two output cells on which Softmax Activation is applied. The cell with a higher probability determines the intent.

### 5.4 Emotion Classifier

The Emotion Classifier also incorporates an LSTM Unit with 300 cells, followed by a Dense layer consisting of 100 units. The major difference from the intent classifier is the number of units in the last Dense layer. As we have already categorized the emotions into nine different groups, the last Dense layer will have nine units. The Softmax Activation is applied to the outputs, and the unit having the maximum probability predicts the emotion.

We used Adam Optimizer and Categorical Cross-Entropy Loss for training both of the classifiers.

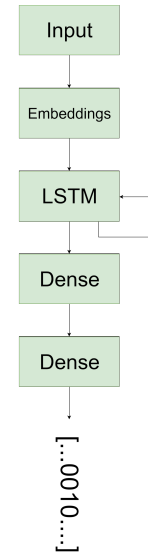


Figure 8: Emotion / Intent Classifier

### 5.5 Response Generation

The medical answer generation architecture consists of two inputs, the utterance, and the response. Initially, the response would be an empty string, and the question is the immediate user input. We pass both of these through a shared embedding layer, which is initialized by the pre-trained glove embedding weights. Two separate LSTM units



consisting of 300 cells are used, one to the utterance embedding and second to the response embedding. We concatenate the outputs of these LSTM cells, and then a dense layer is applied to this concatenated output. The output of this dense layer is fed into the second dense layer. This second dense layer has output dimensions equal to the vocabulary size. After applying Softmax, the output cell having the maximum probability corresponds to a word in the vocabulary.

Response and utterance are the two inputs for the model at each stage. Ground truth word is the output corresponding to the input utterance. While training, we take the utterance to be the question present in the dataset. An empty string is passed as a response initially. Then we calculate the categorical cross-entropy loss for the predicted word. Response fed in the corresponding step is word-by-word incremented with the ground truth value of the previous step. Furthermore, it continues until we reach the end of the sentence. The number of training instances also increases by a lot as we create multiple training instances from a single question answer.

The method of training that we have opted for the text generation is the Teacher Forcing Method (Lamb et al., 2016). Teacher forcing is a method in which we use the output generated by the model in the previous time step to generate the output in the next time step. While testing, we initialize the utterance by the user input. We initialize the response by an empty string. At every instance, the model predicts the next word. We append this word to the response, and the model predicts the next word by using the new response. It continues till we predict the end of sentence tag, or we exceed a pre-defined sentence length.

An empathetic response is also generated using a similar model. For generating an appropriate empathetic response, we need to consider the previous context along with the current user input. The previous input-response pair contains information about the context of the conversation. We have considered the last input-response pair along with the current input to generate the current response. For forcing the response to be empathetic corresponding to the emotion of the user input, we prepended the emotion obtained from the emotion classifier in the current user input. Four separate LSTMs were used, three for the past input,

past response, the current input, and the current response. The concatenate step merges the output of these four LSTMs. The remaining architecture, including two dense layers, remains the same. We initialize the current response by an empty string and follow the similar teacher forcing method we used for training and testing the medical answering model.

## 6 Training

### 6.1 Classifiers

Both of the classifiers are based on LSTMs as shown in **Figure.8**. These classifiers were trained on a batch size of 64. The number epochs that they were trained was 20 around which both of the model saturated.

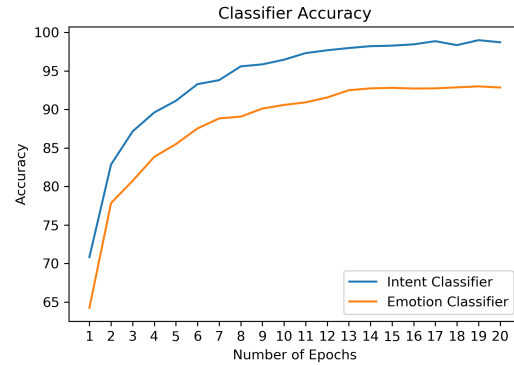


Figure 9: Classifier Accuracy

### 6.2 Text Generators

The conversation generation model is based on parallel LSTMs as shown in **Figure.7**. This conversation generation model was trained with a batch size of 16, which meant, at each training instance the model computed losses out of its predicted conversations and then back-propagated those losses. With the help of regularization methods such as Dropouts and weight decay, the model was trained for a total of 80 epochs at which it got saturated.

## 7 Evaluation

**Figure.9** shows the accuracy of our classifiers. It can be observed from this graph that the *Intent Classifier* saturated at an accuracy of 98.7 percent whereas the *Emotion Classifier* saturated at an accuracy of 92.4 percent.

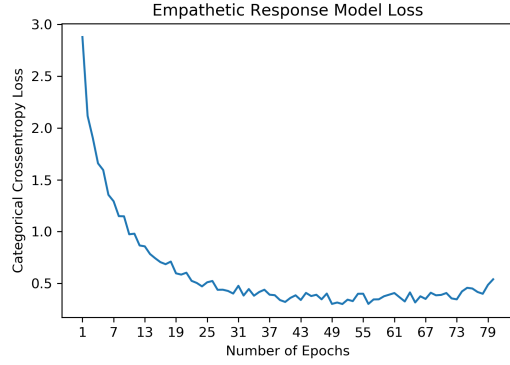


Figure 10: Conversation Generation Loss vs Epochs

For the conversation generation model which is based on parallel LSTMs, **Figure.10** shows its *Categorical Cross Entropy* loss versus the number of epochs it was trained on. It can be seen that the model got saturated at a loss value of 0.48.

We have used automated metrics such as BLEU(Papineni et al., 2002a) score and BERT score for the model response, and compare them against the actual response. BLEU is an n-gram based method to compare the similarity between a candidate text and one or more reference text. (Papineni et al., 2002b)

BERT-Score is an automatic evaluation metric for text generation. Similar to some of the other metrics, it also computes a similarity score. Instead of finding an exact match, the contextual embedding of each token in the candidate sentence is compared with the embeddings of all tokens in the reference sentence. The embeddings are compared based on cosine similarity.(Zhang et al., 2019)

The evaluation of the response generator model are as follows:

	CARO	Facebook AI
<b>BLEU Score</b>	0.179	0.08
<b>BERT Score</b>	0.83	—

Table 2: Evaluation Scores

The emotion and the Intent Classifier accuracy are as follows:

	CARO
<b>Intent Classifier</b>	98.7%
<b>Emotion Classifier</b>	92.4%

Table 3: Classifier Accuracy

Few examples of the utterance-response pairs as generated by our model are shown in the Figure:11.

Utterance	Response
i am lonely	oh that is so sad
i lost my job today	oh sorry to hear that
i won best employee award	congratulations you must have worked hard for that
how do i quit smoking?	quitting is all about being steadfast
how do i stop hairfall?	hairfall can be due to many reasons like stress
how can i lose weight?	regular exercise with more fiber intake will be helpful

Figure 11: Utterance-response pair generated by CARO

## 8 Conclusion

In this study, We have attempted to develop a chatbot emphasizing the issues that are faced by the people with major depression. In this, we addressed many issues with the current implementations of chatbot like monotonic responses and lack of any medical support. The novelty in our approach was to inculcate the generation of medical responses along with maintaining the basic empathetic conversation with the intended user.

For the same, We have developed a novel Neural Architecture based on parallel LSTMs for generation of empathetic responses as well as for the medical responses. We have also shown that considering the last three utterances enables us to produce more appropriate empathetic responses by preserving the ongoing context. Our proposed chatbot also has achieved A BLUE score of 0.179 which is a significant improvement as compared to the one developed by Facebook-AI research. The ability of our chatbot to provide medical responses along with an empathetic response has not only increased the results but also makes it much more reliable and usable for the case of people having major depression.



## References

- Brown et al. Apr 2011. Medication adherence: Who cares? <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3068890/>.
- Fabio Catania, Micol Spitale, and Davide Fisicaro. 2019. Cork: A conversational agent framework exploiting both rational and emotional intelligence. In *IUI Workshops*.
- Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.
- S. Divya, Indumathi, S. Ishwarya, M Priyasankari, and S KalpanaDevi. 2018. A self-diagnosis medical chatbot using artificial intelligence.
- Dongkeon Lee, Kyo-Joong Oh, and Ho-Jin Choi. 2017. The chatbot feels you - a counseling service using emotional response generation. In *2017 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 437–440.
- eHealthForum. 2019. Symptom, side effect, syndrome, surgery, or disease. [https://ehealthforum.com/health/health\\_forums.html](https://ehealthforum.com/health/health_forums.html).
- Gyant. 2016. Ai patient engagement solution <https://gyant.com/>.
- HealthTap. 2019. High-quality, affordable healthcare created for the modern workplace. <https://www.healthtap.com/>.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Simon Hoermann, Kathryn McCabe, David Milne, and Rafael Calvo. 2017. Application of synchronous text-based dialogue systems in mental health interventions: Systematic review. *Journal of Medical Internet Research*, 19:e267.
- Andres Camacho Nicole ROSENDIN David Merkoski Kevin A., Mark Berman.
- Alex Lamb, Anirudh Goyal, Ying Zhang, Saizheng Zhang, Aaron Courville, and Yoshua Bengio. 2016. Professor forcing: A new algorithm for training recurrent networks.
- Jia Li, Xiao Sun, Xing Wei, Changliang Li, and Jianhua Tao. 2019. Reinforcement learning based emotional editing constraint conversation generation.
- Lasse Nielsen. 2017. Medical question answer data <https://github.com/lasseregim/medical-question-answer-data/>. <https://github.com/LasseRegin/medical-question-answer-data/>.
- Mohammad Nuruzzaman and Omar Hussain. 2018. A survey on chatbot implementation in customer service industry through deep neural networks. pages 54–61.
- Endang Wahyu Pamungkas. 2019. Emotionally-aware chatbots: A survey.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002a. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02*, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002b. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02*, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Thomas Polzin and Alex H. Waibel. 2000. Emotion - sensitive human - computer interfaces.
- BMC Psychiatry. May 2017. Associations of fatigue to work-related stress, mental and physical health in an employed community sample. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5420158/>.
- Indian Journal of Psychiatry. March 2017. National mental health survey of india 2015–2016. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5419008/>.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019a. Github - towards empathetic open-domain conversation models: a new benchmark and dataset. <https://github.com/facebookresearch/EmpatheticDialogues>.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019b. Towards empathetic open-domain conversation models: a new benchmark and dataset. In *ACL*.
- Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019c. Towards empathetic open-domain conversation models: a new benchmark and dataset. In *ACL*.
- Ranci Ren, John W Castro, Silvia T Acuña, and Juan de Lara. Usability of chatbots: A systematic mapping study.
- Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2015. Building end-to-end dialogue systems using generative hierarchical neural network models.
- Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks.

Kyle Swanson, Lili Yu, Christopher Fox, Jeremy Wohlwend, and Tao Lei. 2019. [Building a production model for retrieval-based chatbots.](#)

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need.](#)

WebMD. 2019. Health a-z. <https://www.webmd.com/g00/?i10c.ua=1&i10c.encReferrer=aHR0cHM6Ly93d3cuZ29vZ2x1LmNvbS8%3d&i10c.dv=1>.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. [Bertscore: Evaluating text generation with bert.](#)

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2018. [The design and implementation of xiaoice, an empathetic social chatbot.](#)