Mental Healthcare Chatbot Using Sequence-to-Sequence Learning and BiLSTM

Chapter · September 2021

DOI: 10.1007/978-3-030-86993-9_34

Afsana Rakib

North South University

1 PUBLICATION **2** CITATIONS

SEE PROFILE

Md. Monsur Hillas

North South University

2 PUBLICATIONS **3** CITATIONS

SEE PROFILE

Mental Healthcare Chatbot using Sequence-to-Sequence Learning and BiLSTM

Afsana Binte Rakib[1][0000−0002−2001−9484], Esika Arifin Rumky[2][0000−0002−4223−7343], Ananna J Ashraf[3][0000−0002−2895−3896], Md. Monsur Hillas[4][0000−0003−3358−3436], and Muhammad Arifur Rahman[5][0000−0002−6774−0041]

[1] Department of ECE, North South University, Dhaka, Bangladesh
afsana.rakib@northsouth.edu
[2] Department of ECE, North South University, Dhaka, Bangladesh
esika.rumky@northsouth.edu
[3] Department of ECE, North South University, Dhaka, Bangladesh
ananna.ashraf@northsouth.edu
[4] Department of ECE, North South University, Dhaka, Bangladesh
monsur.hillas@northsouth.edu
[5] Department of Physics, Jahangirnagar University, Dhaka, Bangladesh
arif@juniv.edu

Abstract. Mental health is an important aspect of an individual's well-being which still continues to remain unaddressed. With the rise of the COVID-19 pandemic, mental health has far continued to decline, especially amongst the younger generation. The aim of this research is to raise awareness about mental health while simultaneously working towards removing the societal stigma surrounding it. Thus, in this paper, we have created an integrated chatbot that is specifically geared towards mentally ill individuals. The chatbot responds empathetically which is built using a Sequence-to-Sequence (Seq2Seq) encoder-decoder architecture. The encoder uses Bi-directional Long Short Term Memory (BiLSTM). To compare the performance, we used Beam Search and Greedy Search. We found Beam Search decoder performs much better, providing empathetic responses to the user with greater precision in terms of BLEU score.

Keywords: Depression · Chatbot · BiLSTM · RNN · Encoder · Decoder · Greedy · Beam

1 Introduction

Mental health issues have continued to prevail among the population for decades. According to [11], the rate of annual self-harm by girls was found to be 37.4 and 12.3 in boys. However, most of the participants have no record of following up with mental health services. Although this study was carried out in a limited scope, it sheds light on the need for immediate intervention to help and support the myriads of people suffering in silence in a manner that is both sensible and affordable. Furthermore, studies have shown that there is an increase in Artificial Intelligence (AI) based chatbots to promote mental well-being, among which Wysa is one such example where a high percentage of the users have shown to develop improved moods [7]. Therefore, it is evident that AI-based conversational agents can work towards the improvement of overall mental well-being. This is where our chatbot comes into play. Our bot integrates Machine Learning and executes it in a practical manner by communicating very closely with the users through a specialized algorithm. Through a Sequence-to-Sequence Learning mechanism, our model has a conversation with the user by generating appropriate responses. To evaluate the performance, we have a conversation with the bot using sentences as user inputs and calculate the BLEU score for each generated model, which uses Beam and Greedy Decoder architectures, respectively. The contribution of the paper is as follows:

- Treating mental health is costly and societal stigma bars individuals from seeking help.
- Mentally ill individuals will be benefitted from this chatbot as it will provide empathetic responses whenever they feel the need for someone to listen.
- For future improvements, we can use the data of the conversation to diagnose a patient with any disorder they may have.

2 Literature review

According to [2], the most common barriers when it comes to seeking mental health services were attitudinal, lack of finance, and availability. This raises an immediate need for a system that acts as a supplementary support system in place of healthcare professionals in the field. This is where artificial-enabled conversational technologies come into play as an empathetic virtual support system. This is evident in the study conducted by [19] which reveals that a person discloses deeper emotions from the conversations with real-time mental healthcare chatbots. This is because of the fact that the users do not receive any judgmental attitudes which can be usually expected in the case of human-to-human interactions.

Authors Prakash et. al collaborated to build a chatterbot using an RNN Encoder–Decoder composed of two RNNs. The proposed model of the chatbot is implemented by using the Sequence-To-Sequence (Seq2Seq) model with transfer learning [20]. The fixed-size context vector generated by the encoder is given as the input to the decoder of RNN. For the dataset, the model uses a movie dialog corpus of 220,579 conversational exchanges. Long Short Term Memory (LSTM) consists of two activation functions- *Sigmoid* and *tanh*. The *Sigmoid* function generates the correct details of the memory with the new input but cannot remove the memory. The *tanh* function controls the values across the network by which conversations go on. LSTM based encoder-decoder structure has been demonstrated to be more vigorous, cleaner, and quicker than an Artificial Neural Network (ANN) model. Moreover, this model has shown to generate further accurate grammar as compared to other generative models.

Yin et. al proposed a deep learning chatbot, Evebot, that uses Seq2Seq model through LSTM in order to diagnose negative user emotions and prevent depression. The proposed system combines BiLSTM network, Seq2Seq neural network, and Maximum Mutual Information (MMI) model [25]. This system was primarily targeted towards campus students as a means of virtual psychological therapy. The chatbot using the sentiment analysis model analyzes the user emotion throughout the conversation while the classifier model is responsible for the responses of the chatbot. The BiLSTM model was used to classify the user responses into two categories - positive and negative. The BiLSTM model provided a 90.91% precision of the test set. The Seq2Seq model was built on the MMI method which analyzes the mutual dependence between the input and output. However, this system cannot remember the previous topic of conversation and generate different responses to questions of the same topic. This may confuse the users. To solve this problem, some elements of a rule-based chatbot could be added.

3 Methodology

Our model is based on a Seq2Seq encoder-decoder architecture. The encoder uses Bi-directional Long Short Term Memory (BiLSTM) units which is a variation of LSTM networks. LSTM is an extension of Recurrent Neural Network (RNN) which can control the overall flow of information within the network. LSTM primarily helps in preserving the error so that it can be dealt with through back-propagation in different time and layers [13], [12], [23]. LSTM contains an input gate, an output gate and a forget gate [24]. The gates have an additional layer by the name of the Sigmoid neural network layer. Since the gates control the flow of information, the Sigmoid layer is crucial as it produces outputs between the numbers zero and one, where zero represents "letting no information

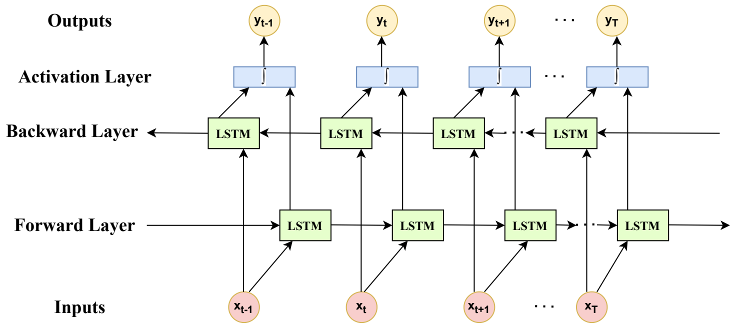

Fig. 1. BiLSTM Architecture: BiLSTM networks have two side-by-side layers of LSTMs where the actual input sequence is given as an input to the first layer and the reversed version of the input sequence is provided as an input to the second layer [17]. BiLSTMs, therefore, essentially increase the quantity of information that is available to the network which further improves the contexts of the algorithm.

through" while one represents "letting all the information pass" [14]. The input gate chooses the values to be updated and the tanh layer generates a vector of new candidate values which are then added as additional information to the cell state. However, in the model used, BiLSTM is considered for making fast predictions in real-time [16]. It has internal memory and can use the input to predict the next state. Since the chatbot is reliant on sequential data such as speech, it is helpful to develop an upgraded model. The encoder uses BiLSTM units which generates better results than unidirectional LSTM as it considers the embeddings of both the previous words and the next words suitable for predicting the target variable [26]. Furthermore, since the general Seq2Seq decoder has a possibility of information loss, we use the attention mechanism as proposed in [3] where the output is not solely dependent on the encoder's context vector. Rather, the decoder also pays attention to certain parts of the input. The attention is computed by taking the encoder's outputs and the current hidden state of the decoder. Here, the shape of the output attention weights matches that of the input sentence; hence, we can multiply them by the encoder outputs. This generates a weighted sum that guides the decoder to pay attention to the parts of the encoder's output. The BiLSTM architecture is shown in Figure 1.

The Sequence-to-Sequence model with attention mechanism is given in Figure 2. A typical Sequence-to-Sequence (Seq2Seq) model consists of the following three components [15]:

i) Embedding - The embedding layer converts the input sentence into a vector of numbers.

ii) Encoder – The encoder has BiLSTM units. The embeddings of inputs - variable-length vectors - are processed to convert into fixed-length vectors.

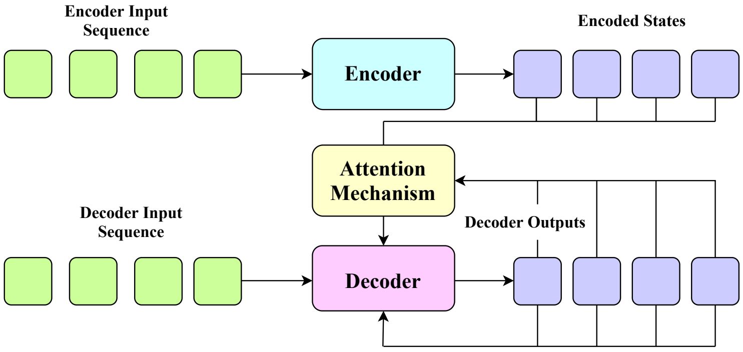

Fig. 2. Sequence-to-Sequence Model with Attention Mechanism: The encoder input sequence is passed into the encoder to receive the encoded states. Then, the encoded states from each time step are forwarded to the attention mechanism in order to choose the important encodings needed for the decoder [8]. During decoding, both the decoder and encoder states are passed into the feed-forward network which returns weights for each of the encoder states.

iii) Decoder – The decoder has a Gated Recurrent Unit (GRU) network which is a variation of LSTM. GRU merges both the forget and input gates of the network into one single gate known as the “update gate” [6].

As our dataset consists of sentences that are sequences of words, we mapped each unique word to an index value using a vocabulary class. This class adds words to the vocabulary and counts the frequency of the words. The questions and answers from the dataset used functions as sentence pairs. To reduce the complexity of the function and to achieve faster convergence, we trim rare words which are repeated infrequently and filter out the sentences containing these rare words. The non-letter characters are also removed and the words are formed into tokens. After mapping, the encodings are then given as input to the neural network. To aid in convergence while training, the sentences having a length of greater than 1000 words are filtered out.

At each time step, we manually feed the input batch to the decoder. Based on the decoder’s tensors, we calculate the masked loss. The average negative log-likelihood is calculated for those elements that correspond to a 1 in the mask tensor. For a single training iteration, a single batch of inputs is trained. RNNs are usually prone to vanishing gradient [9], [10]. However, since we used LSTM, we implement gradient clipping to solve the vanishing gradient problem, thus, we prevent the gradients from growing exponentially. This also aids in convergence. We also use teacher forcing to aid in further efficient training. Here, the current target word is used as the next input of the decoder rather than using the decoder’s current prediction. This model uses softmax function as its activation function:

$$p_{\mathrm{i}} = \frac{e^{\mathrm{a_i}}}{\sum_{k=1}^{N} e_k^a} \tag{1}$$

where p_i is a layer and a is the input vector to the i^{th} layer; $a = [a_1, a_2 \ldots.., a_N]$.

For the evaluation of a string input, we created a user interface for the chatbot where the user can communicate with the chatbot through text. After pressing Enter on the keyboard, the text is fed to the model to obtain a decoded output sentence. We can press “q” or “quit” to stop chatting. Finally, if a word in the sentence is not found in the vocabulary, an error message is generated and the user is prompted to enter a different sentence.

4 Data Set

We have used The Mental Health FAQ by [18] which consists of 98 questions and answers related to mental health. Each of the questions is provided with a unique ID and the aim is to help people with mental health problems. The dataset was compiled taking information from the Kim Foundation, Mental Health America, Wellness in Mind, and heretohelp organizations. This dataset was collected from Kaggle.

The other dataset is solely dedicated to creating a model showing the conversation between a client and a therapist based on grounded theory analysis [22]. The main idea proposed here is a model of the therapist’s inner conversation consisting of four positions. Using this idea, the conversation between the therapist and the client is conducted and we have incorporated these conversations as the dataset for training. The dataset mostly functions as a platform where the therapist performs their task in treating the client (patient) by listening to the client's story. Based on the analysis, the therapist responds to the client. As such, our chatbot is designed to function in a similar manner, opting to listen to the user vent about their innermost fears and issues.

5 Results and Analysis

During the preprocessing of data, we set a threshold of maximum length of sentences and filtered out which were longer than the threshold in order to aid in training convergence. The learning rate is kept 0.0001. To achieve faster convergence during the training process, we trimmed the least used words out of the vocabulary. Here, the current target word is used as the next input of the decoder rather than using the decoder´s current prediction. The model uses the softmax as its activation function and it is used in parallel with the cross-entropy loss function. The paper compares the models developed with Beam Search decoder and Greedy Search decoder respectively while other conditions such as the learning rates, number of iterations, teacher forcing ratio, and the data sets are kept constant. With Greedy search, the algorithm selects the most frequently appearing words at each time-step in the output sequence [5]. On the other hand, Beam search is an expansion of the Greedy search which returns the output sequences which are most likely to occur.

Table 1. An example of a conversation between the user and the chatbot using the Beam decoder model.

User	**Chatbot**
Hi	hello..
I feel like a loser , what do I do?	many people find peer support a helpful tool that can aid in their recovery . there are variety of organizations that that offer support groups for their family members members members members and and support groups are peer led while led by by by mental mental health health health health
Why do not my parents talk to me?	i don good talk . mentioned it it they do to to to .

The performance of our chatbot model can be further improved and the slow training rate can be overcome given larger data sets and GPU architecture. We further noticed that with increasing iterations, the chatbot's performance increases. This was expected as training a neural network model typically requires many iterations. The downside is that running the model is more time-consuming which is not efficient in the long run. Tables 1 and 2 show the results obtained

Table 2. An example of a conversation between the user and the chatbot using the Greedy decoder model.

User	**Chatbot**
Hi	..
I feel like a loser , what do I do?	many many many peer support their their . a variety of consumers their and may may may may may may may may may . . professional .
Why do not my parents talk to me?	now that you mentioned it it they they do to me .

from chatting with the bot using a model with Beam Search Decoder and Greedy Search Decoder respectively.

A set of questions were formed based on the data sets used and both the models were presented with the same questions. As expected, it was observed that the model using the Beam decoder provided better responses than the one with the Greedy decoder. However, the Greedy decoder is faster in terms of responses as it takes less time to respond.

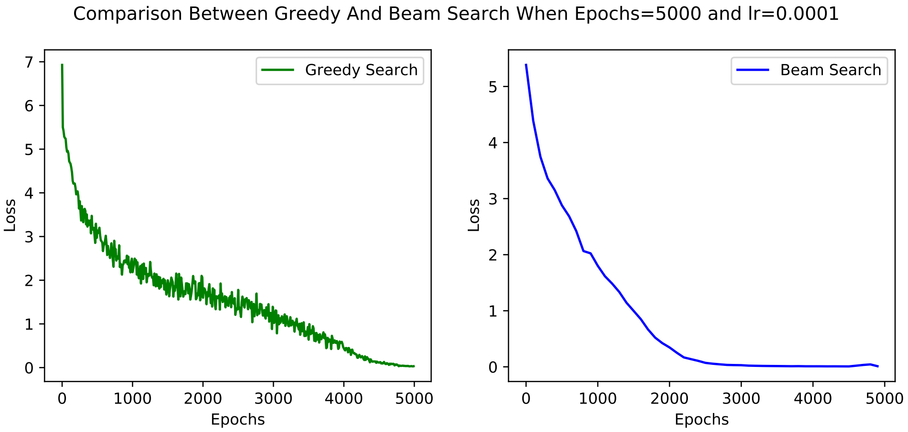

Fig. 3. It is observed that the curve for the Greedy model has a lot of spikes with a gentler slope. There is a considerable reduction in loss values with increasing iterations. The curve for the Beam model has comparatively fewer fluctuations; hence, it is smoother. However, it is slightly steeper than the other one. In both the models, the learning rate is kept 0.0001 with 5000 epochs.

Figure 3 shows the loss values with increasing iterations during the training process. In both the curves, loss values decrease significantly with an increasing number of iterations. This is expected since the loss should decrease as the predicted probability converges to the actual label. The model using the Greedy search decoder runs faster than the Beam one. Hence, it can be observed that as we increase the number of iterations in tandem with the precision of the learning rate, the model will learn better and be more efficient in the long run and the errors will also decrease considerably.

We further calculated the individual Bilingual Evaluation Understudy (BLEU) score for both models. The BLEU score is a measure of how similar the candidate text and the reference texts are. This score ranges from 0.0 to 1.0 with values closer to 0.0 indicating less precision and values closer to 1.0 indicating greater precision [4]. The BLEU score for 1-gram on testing pairs and 2-gram on testing pairs were generated for each of the models where 1-gram represents a single word and 2-gram represents word pairs. Figure 4 shows the BLEU score comparison between the Beam search and the Greedy search models.

Although both models provide a good level of precision, the model implementing the Beam search decoder generates a better precision than the Greedy one for 1-gram while they perform with a similar precision for 2-gram on testing pairs. Moreover, even though Greedy is fast, the model with the Beam search decoder provides relatively optimal responses than the Greedy one when evaluating the quality of the final output received from the conversation with the chatbot. Therefore, Beam is a heuristic search algorithm that performs better than the Greedy model. Yet, we believe that there might be a slight deviation of the results as shown by [1] if we select a different set of data set or a non parametric model setup like Gaussian process [21].

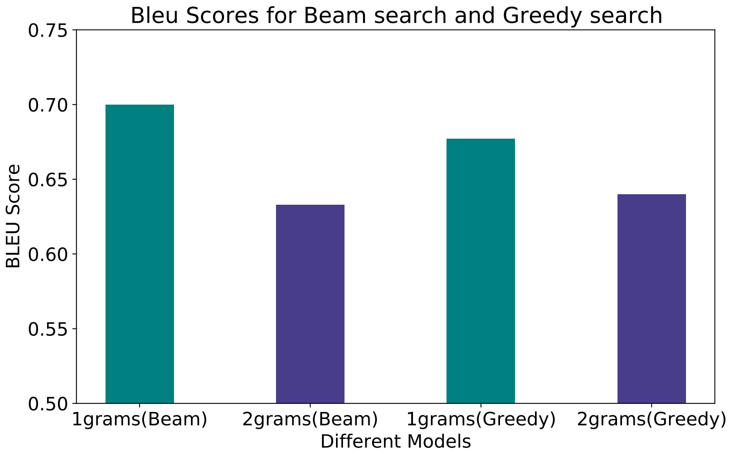

Fig. 4. For the model with the Beam search decoder, the BLEU score for 1-gram on testing pairs was 0.700 and the score for 2-gram on testing pairs was 0.634. For the model with the Greedy search decoder, the BLEU score for 1-gram on testing pairs was 0.678 and the score for 2-gram on testing pairs was found to be 0.640.

6 Conclusion

One of the main takeaways of this paper is for everyone to be aware of mental health as a whole more and to put importance into it, rather than casting it aside. In a society where mental health is considered taboo, perhaps an artificially inspired brain is a better listener than an actual person. In this paper, we built a chatbot that uses a Sequence-to-Sequence (Seq2Seq) encoder-decoder architecture where the encoder uses BiLSTM. Then we compared the performance of the model; we used two decoder architectures - one using Beam Search and the other using Greedy Search. Between two of the chatbot models, the model using the Beam search decoder performs much better with empathetic responses, which is the primary aim of the chatbot. However, it should be noted that it is by no means the end of advancement. Through further training, it can be made more empathetic to cover various emotional scenarios. The model has gained considerable precision and will continue to do so as we expand the dataset and train it further. In addition to that, we can create separate logs for each user to be of use therapeutically. Through daily human interactions, it can learn more and increase its precision through reinforcement learning.

References

1. Adiba, F.I., Islam, T., Kaiser, M.S., Mahmud, M., Rahman, M.A.: Effect of corpora on classification of fake news using naive bayes classifier. International Journal of Automation, AI and Machine Learning, Canada **1**, 80–92 (2020)
2. Andrade, L.H., Alonso, J.: Barriers to mental health treatment: Results from the who world mental health (wmh) surveys. Psychological Medicine **44**(06), 15 (August 2013). https://doi.org/10.1017/S0033291713001943
3. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv **1409**, 15 (September 2014)
4. Brownlee, J.: A gentle introduction to calculating the bleu score for text in python, https://machinelearningmastery.com/calculate-bleu-score-for-text-python/
5. Brownlee, J.: How to implement a beam search decoder for natural language processing, https://machinelearningmastery.com/beam-search-decoder-natural-language-processing/
6. Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv p. 15 (June 2014). https://doi.org/10.3115/v1/D14-1179
7. Inkster, B., Sarda, S., Subramanian, V.: A real-world mixed methods data evaluation of an empathy-driven, conversational artificial intelligence agent for digital mental wellbeing. JMIR mHealth and uHealth **6**, 14 (September 2018). https://doi.org/10.2196/12106
8. Lintz, N.: Sequence modeling with neural networks (part 2): Attention models, https://indico.io/blog/sequence-modeling-neural-networks-part2-attention-models/
9. Mahmud, M., Kaiser, M.S., McGinnity, T.M., Hussain, A.: Deep learning in mining biological data. Cognitive Computation **13**, 33 (January 2021). https://doi.org/10.1007/s12559-020-09773-x

10. Mahmud, M., Kaiser, M.S., Rahman, M.M., Rahman, M.A., Shabut, A., Al-Mamun, S., Hussain, A.: A brain-inspired trust management model to assure security in a cloud based iot framework for neuroscience applications. Cognitive Computation **10**, 864–873 (2018)
11. Morgan, C., Webb, R.T.: Incidence, clinical management, and mortality risk following self harm among children and adolescents: cohort study in primary care. BMJ Clinical Research **359**, 9 (October 2017). https://doi.org/10.1136/bmj.j4351
12. Nasrin, F., Ahmed, N.I., Rahman, M.A.: Auditory attention state decoding for the quiet and hypothetical environment: A comparison between blstm and svm. Proceedings of TCCE, Advances in Intelligent Systems and Computing (2020)
13. Noor, M.B.T., Zenia, N.Z., Kaiser, M.S., Mamun1, S.A., Mahmud, M.: Application of deep learning in detecting neurological disorders from magnetic resonance images: a survey on the detection of alzheimer's disease, parkinson's disease and schizophrenia. Brain Informatics **7**, 11 (October 2020). https://doi.org/10.1186/s40708-020-00112-2
14. Olah, C.: Understanding lstm networks, http://colah.github.io/posts/2015-08-Understanding-LSTMs/
15. Palasundram, K., Sharef, N.M., Nasharuddin, N.A., Kasmiran, K.A., Azman, A.: Sequence to sequence model performance for education chatbot. International Journal of Emerging Technologies in Learning (iJET) **14**(24), 56 (December 2019). https://doi.org/10.3991/ijet.v14i24.12187
16. Papers with Code: Bidirectional LSTM, https://paperswithcode.com/method/bilstm
17. Papers with Code: Long short-term memory, https://paperswithcode.com/method/lstm
18. Prabhavalkar, N.: Mental health faq, https://www.kaggle.com/narendrageek/mental-health-faq-for-chatbot
19. Prakash, A.V., Das, S.: Intelligent conversational agents in mental healthcare services: A thematic analysis of user perceptions. Pacific Asia Journal of the Association for Information Systems **12**, 34 (June 2020). https://doi.org/10.17705/1pais.12201
20. Prakash, K.B., Nagapawan, Y., Kalyani, N.L., Kumar, V.P.: Chatterbot implementation using transfer learning and lstm encoder-decoder architecture. International Journal of Emerging Trends in Engineering Research **8**, 7 (May 2020). https://doi.org/10.30534/ijeter/2020/35852020
21. Rahman, M.A.: Gaussian Process in Computational Biology: Covariance Functions for Transcriptomics. Ph.D. thesis, University of Sheffield (2018)
22. Rober, P., Ellliott, R., Buysse, A., Loots, G., Corte, K.D.: Positioning in the therapist's inner conversation: A dialogical model based on a grounded theory analysis of therapist reflections. Journal of Marital and Family Therapy **34**(3), 16 (August 2008). https://doi.org/10.1111/j.1752-0606.2008.00080.x
23. Sadik, R., Reza, M.L., Noman, A.A., Mamun, S.A., Kaiser, M.S., Rahman, M.A.: Covid-19 pandemic: A comparative prediction using machine learning. International Journal of Automation, AI and Machine Learning, Canada **1**, 1–16 (2020)
24. Sak, H., Senior, A., Beaufays, F.: Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech p. 5 (February 2014)
25. Yin, J., Chen, Z., Zhou, K., Yu, C.: A deep learning based chatbot for campus psychological therapy. arXiv **8**, 31 (October 2019)
26. Yin, W., Kann, K., Yu, M., Schütze, H.: Comparative study of cnn and rnn for natural language processing. arXiv p. 7 (February 2017)