Towards an Emotion Based Chatbot System to Tone Down Mental Health Problems*

Mrulay Mistry
University of Windsor
mistry13@uwindsor.ca

Sarah Youssef
University of Windsor
youssefs@uwindsor.ca

Ashutosh Sadana
University of Windsor
sadana@uwindsor.ca

Mudita Sharma
University of Windsor
sharma7s@uwindsor.ca

ABSTRACT

It is normal among people to be conservative about their mental health issues and never consult a specialist, although such issues can remain unnoticed to the person himself until severe physical and (or) psychological disorders are caused. In this paper, we introduce the core block of a chatbot system that aims to mitigate people's sufferings due to mental health problems. This system predicts the user's emotional state, generates the appropriate context-based replies that comfort them, and suggests activities based on the detected emotion(s). We have trained our system with the CARER dataset for 6 different labeled emotions: sadness, joy, love, anger, fear and surprise. Among the tested algorithms, Linear Support Vector Classifier (LSVC) algorithm showed the highest F1 score with 78% when evaluated with the ISAER dataset during the emotion detection part. Moreover, the emotion-aware text generation model was 87% accurate when tested with the Cornell Corpus database, corresponding to a 0.30 cross-entropy loss.

KEYWORDS

Artificial Intelligence, Mental Health, Emotion Detection, Text Generation, Activity Suggestion

1 INTRODUCTION

Mental health is a serious problem for humans, especially teenagers. Nonetheless, most patients find difficulties in sharing their emotions, even with psychologists. Holding up to these feelings might result in radical behavioral changes affecting people's mental hygiene. Moreover, it is still shameful in some parts of the world to seek mental professional help, to avoid being labeled as mentally sick by society. Consequently, they do not have anyone's support and keep suffering in silence [1]. As a result, it is important to track the mental well-being of people across the globe using a system that detects users' emotions and tones them down.

To the best of our knowledge, there has not been an automated system for emotional support commercially available, despite the existing research efforts to build such a system. In this article, we propose a system that automatically analyzes the user's emotions, labels them and generates sentences that would be appropriate to comfort the user. The proposed system can also recommend a set of activities to help the users get out of the grief they are suffering. The entirety of the chat-bot system is divided in two phases; The Back-end AI system and the Front end UI. Our paper introduces a novel architecture that deals with machine learning and deep learning for the back-end processing of our proposed system.

The rest of this article is sectioned as follows: Section 2 for literature review, Section 3 for problem statement, Section 4 for methodology, Section 5 for experimental setup, Section 6 for results, Section 7 for discussion and challenges and Section 8 for conclusion and future work.

2 LITERATURE REVIEW

In the past, there have been many works that include detection of emotions either using a single mode[2] of input or multi-modal inputs [3]. However, there is a lot less literature on emotion aware text generation or computer based emotional state alleviation.

Kenneth Mark Colby [4], in 1975 designed a chatbot to behave like paranoid person using a rule-based approach and that was the first time in the history where a computer was made to mimic human emotions. In more recent times, Zhou, et al. [2] in 2014 developed Microsoft Xiaolce, a chatbot that incorporated emotional satisfaction to satisfy the human need for communication. This system considered factors like Emotional Quotient (EQ) and Intelligence Quotient (IQ) to achieve this task. They published an improvement in this in 2018 [5] where the system was able to generate context and dig out implicit emotions states. Zhang et al. [6] proposed a system which produces several responses for every emotion category in their dataset using Bi-LSTM model. They categorized the emotions in mainly 5 categories. Hu et al. [3] proposed the detection of emotions using audio inputs. Instead of a lexicon based approach, they used the tone to determine the intensity of emotion of the user.

3 PROBLEM STATEMENT

Although previous researchers have implemented emotion-detection systems by using either a single-mode input or multi-modal inputs, most of these systems did not focus on emotion-aware text generation for the users. We are proposing a system that not only detects the emotion conveyed by the users, but also generates replies and suggests activities to tone down their mental health problems. This project aims to free the users from their emotional suffering. The system must be capable of: (i) classifying the user's state of mind; (ii) communicating considerate and relaxing sentences to the user; and (iii) proposing comforting activities based on the detected emotional category. The system will deploy the concept of deep neural networks in artificial intelligence. We will be referring to the OCC model and the Core-Affect model.

*https://github.com/Mrulay/EmotionAwareChatbot

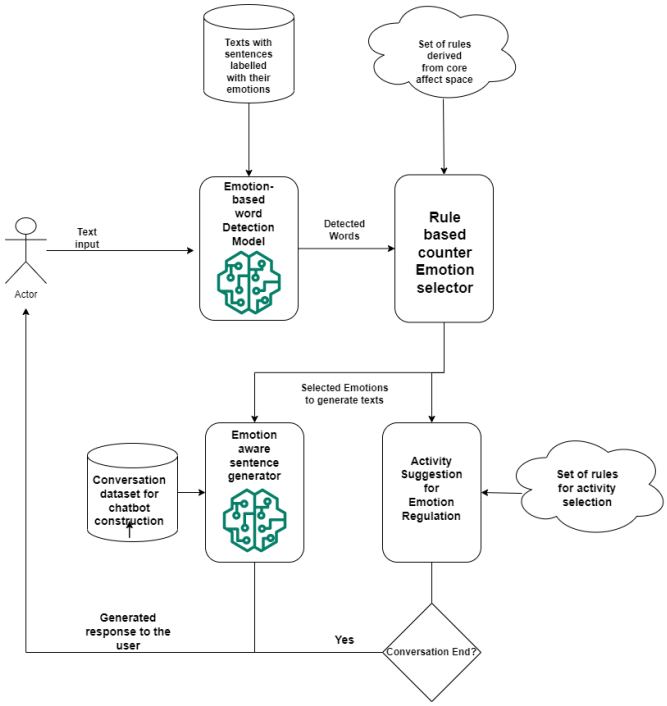

Figure 1: Design Architecture of the Chatbot Backend

4 METHODOLOGY

4.1 Design Architecture

The proposed system is divided into three parts: *Emotion Detection*, *Emotion Aware Reply Generation* and *Activity Suggestion*. The first two parts are two classical challenges for systems in the domain of *Natural Language Processing*. In the first part, *Emotion Detection*, the system receives a text from the user, analyzes it to predict the emotions, and forwards them to the next block. In the second part, *Emotion Aware Reply Generation*, the system generates appropriate texts that are relevant to the forwarded emotions from previous parts. The generated texts are then displayed to the user based on the interactive conversation between them and the bot. The last part, *Activity Suggestion*, is triggered when the interactive conversation ends. Depending on the detected emotional condition(s), the system will recommend a set of activities that would help the patient move to a better state of mind. Figure 1. Shows an overview of how the entire system is designed.

4.2 Emotion Detection

In the domain of natural language processing, text classification is a well explored problem. The Emotion Detection is also a multi-class text classification problem as we simply classify subject sentence in one of the many categories of different emotions.

For this purpose, we chose the CARER dataset[7]. This dataset includes 16000 sentences each labelled with an emotion from a set of emotions *(sadness, joy, love, anger, fear and surprise)*. To prepare the data for modelling, we passed the texts through a pre-processing pipeline that includes removal of 'stopwords', removal of punctuation marks, expanding all the contractions and converting all texts to lowercase.

Definition 4.1. **Stop Words** Sometimes, some extremely common words which would appear to be of little value in helping select documents matching a user need are excluded from the vocabulary entirely. These words are called stop words.[8]

After the pre-processing step, the dataset was converted to a TFIDF matrix so that the most important words are weighted more.

Definition 4.2. **TF-IDF Values** TF–IDF, short for term frequency inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.

This matrix was then divided into training and test sets and trained on multiple classifiers. Out of all the classifiers we used for training, the *Linear Support Vector Classifier* yielded the maximum accuracy. Thus, it was chosen as our emotion detection classifier.

4.3 Emotion Aware Reply Generation

Text Generation is one of the more challenging problems of natural language processing and understanding. While there are many models and algorithms that are either lexicon based or word frequency based, [9], they do not take in account the delivery of emotion along with the sentence. A deep learning model called *Sequence-to-Sequence (seq2seq)*[10] model is the gold standard model for the text generation task.

Typically, a seq2seq model architecture comprises of 2 neural networks, namely and encoder and a decoder. The encoder takes in a question (also known as the context vector) which is then embedded to vectors (Digitized words) as its input. Each neuron in the encoder is an LSTM cell that process an instance of each word in the sentence and passess it's output to it's neighbouring cell. The output of the last cell of the encoder is called the encoder state (i.e. what the encoder learnt from the input sentence). The decoder is feeded another sentence that is known as the answer sentence along with the encoder states and learns the connections between the context vector and the answer vector. These states are then forwarded to a Softmax layer which then outputs the learned values.

With the seq2seq model, the problem of relevance in the answer gets solved. To make the network learn the implicit emotions of the input sentences, we propose a modified version of seq2seq model. As shown in Figure 2, we first embed the model with VAD embeddings.

VAD embeddings: According to the work proposed by [11], each word has a distinct Valence, Affect and Dominance value. The valence (V) value determines the sentiment of each word (i.e., the higher the valence value, the higher the positivity of the word).

To incorporate the valence value each word in our network, first we convert each word of the corpus to integer values (commonly known as vectors), and then append the V value of each word to those embeddings. We call this new embedding as VAD embeddings. This way, we ensure that the emotion values of the context vector are passed into the network and the encoder and decoder learns from those values.

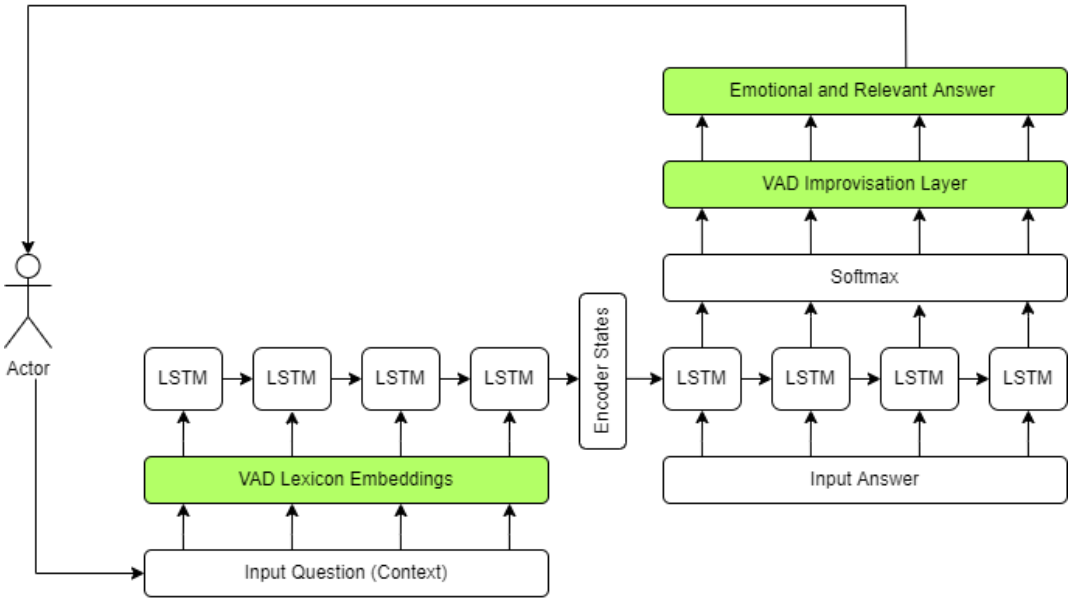

Figure 2: Proposed Network Architecture

At the output layer, after we get the output from the softmax layer, we add one more layer known as the VAD improvisation layer. This layer gets the average V value of the entire output statement, then uses the TFIDF values of each word to get the most important words in the statement and then replaces those words with their synonyms that have a higher V value. This way, we ensure that the resultant statement generated is not harmful for the mental health of the user while it also tries to alleviate the mental state of the user.

To train this model, we used the Cornell Movie Dialog Corpus[12]. This dataset has 220,579 conversational exchanges between 10,292 pairs of movie characters. We chose this dataset as it includes a wide variety of emotions and a large number of vocabulary to go with it.

4.4 Activity Suggestion

After the user ends the conversation with the chat-bot, the system recommends an activity to the user to uplift their emotional state. For every detected emotion, we will select an emotion that we want to induce into the user. For this, we will look into a set of rules that we define ourselves and find the desired emotion that we want. For example, if the user is detected to be sad, the recommended activity would be selected in such a way that would reduce the sadness.

The set of rules are based on the Core Affect Space[1].

The output from the emotion detector will also be passed to the activity suggestion tool. This tool will have a set of activities with a set of weights. I.e., for each activity A, there will be a set of weights $W = (w_1, w_2, w_3, ..., w_n)$. Each weight w_i will represent an emotion and the value of these weights will determine how much an activity induces a particular emotion. Depending on the current emotions of the user, the system will select an activity that maximizes the emotion that we want the user to feel. In other words, the activity with maximum desired weight will be chosen.

5 EXPERIMENTAL SETUP

Since we are using two distinct models in our system, this section consists of two subsections for Experimentation and Evaluation of each model.

5.1 Emotion Detection

We tried 5 different models to train on the CARER dataset. We trained Google's BERT[13], and a custom Recurrent Neural Network with 2 Million parameters and Word2Vec as it's embeddings. Since this is a classification task, we chose the F-1 Score as our testing metric as it provides a good measure of both, the precision and recall. Because of the high score achieved by the Linear Support Vector Classifier (LSVC), we chose this as our main classifier. To test our emotion detection methodology, we applied our trained model on 2 other datasets. (i) **ISAER**[14] and (ii) **Crowdflower Emotion Dataset**[2]

The ISAER contains 3000 documents that reported situations in which emotions were experienced. The Crowdflower Emotion Dataset contains 40,000 tweets scraped from twitter that are categorized in 14 different types of emotions. To feed these datasets with to our model, we only selected the sentences with the following emotion labels: *(sadness, joy, love, anger, fear and surprise)*

5.2 Emotion Aware Reply Generation

This phase of our project can be correlated with the Text Generation problem in Natural Language Processing. To measure the success a

[1]https://illis.se/en/wp-content/uploads/sites/2/2016/11/Core-Affect-Space-written-summary.pdf

[2]https://data.world/crowdflower/sentiment-analysis-in-text

text generator system, we look at the training and testing accuracy and the categorical crossentropy loss. Typically, the Bilingual Evaluation Understudy (BLEU) Score[3] is used for quantifying a sequence to sequence model, but it is only applicable to machine translation models. So, we chose not to go ahead with it.
We chose the **DailyDialog** Corpus[15] and the **AmbigQA** dataset[16]. The DailyDialog dataset contains a total of 13,000 manually labelled conversations with topics and emotions. On the other hand, the AmbigQA covering 14,042 answers that are paired with ambiguous questions.
Hardware Specifications: The experiments on the models described above were run on a computer with an Intel i7 8750H processor running at 2900 MHz using 16 GB of RAM, and an Nvidia Geforce GTX 1060 GPU with 6 GBs of VRAM running on Windows 10.

6 RESULTS

Table 1 summarizes all the F1-Scores of the models that we trained on the datasets mentioned above for emotion detection.

Model	CARER	ISAER	Crowdflower
RNN-W2V	0.60	0.68	0.48
BERT	0.48	untrainable	0.54
Logistic Regression	0.53	0.77	0.61
Multinomial Naive Bayes	0.74	0.74	**0.66**
LSVC	**0.88**	**0.78**	0.63

Table 1: Comparison of F1 Scores for all the tested models on different datasets

As seen in the table, LSVC significantly outperforms all other tested models except on the Crowdflower dataset. Such a discrepancy can be considered as a margin of error.
Table 2 summarizes the accuracies and the losses of our Emotion Aware Reply Generation System.

Model	Cornell	DailyDialog	AmbigQA
Test Accuracy	0.87	0.75	0.61
Crossentropy Loss	0.30	0.21	0.31

Table 2: Accuracies of all tested models for Text Generation

The seq2seq network that we trained produces promising accuracy on the original dataset that we trained on. When we test our model on different datasets, the model proves to be viable as well. A crossentropy score less than 0.2 is said to be acceptable for tasks like text generation and machine translation. As a result, we can say that our model predicts acceptable sentences with the VAD word embeddings. The higher loss and lower accuracy in the AmbigQA dataset can be explained as each question in the dataset is paired with multiple answers. This causes in lowering the probability of the next predicted word, thus, significantly increasing the perplexity of the model.

[3]https://en.wikipedia.org/wiki/BLEU

7 DISCUSSION AND CHALLENGES

Some of the problems that we faced during the development are summarized in this section:
Our aim was to train a perfect text generation model for this system, but training a Seq2Seq model is very expensive. As the Cornell Movie Dialogs Corpus is too huge to train on our system, we had to down sample the dataset size to only 30,000 sentences from 220,000 sentences.
Due to the hardware constraints, the time taken for model training was taking too long. That is why we were not able to reach an optimal cross-entropy loss. Ideally, the cross-entropy loss should be less than 0.2 for the probabilities of a seq2seq model to be considered good.

8 CONCLUSION AND FUTURE WORK

The area of using Artificial Intelligence for the benefit of the mental health of people is an area that is very less explored and our work tries to be a stepping stone in this.
The module for Activity Suggestion was based on our assumptions, if we want to take this further, then verification of the activity weights by a psychoanalyst or a crowdsourced survey followed by statistical significance tests can be done.
Another future improvement in the model would be to use the VAD lexicon corpus to figure out more emotion categories instead of only six in our model. The VAD emotion model[17] paves way for innumerable number of emotions and if used in this paradigm, it can prove to be very beneficial for the task that we are targetting.

REFERENCES

[1] Claire Henderson, Sara Evans-Lacko, and Graham Thornicroft. Mental illness stigma, help seeking, and public health programs. *American journal of public health*, 103(5):777–780, 2013.
[2] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46 (1):53–93, 2020.
[3] Tianran Hu, Anbang Xu, Zhe Liu, Quanzeng You, Yufan Guo, Vibha Sinha, Jiebo Luo, and Rama Akkiraju. Touch your heart: A tone-aware chatbot for customer care on social media. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–12, 2018.
[4] Kenneth Mark Colby. *Artificial paranoia: A computer simulation of paranoid processes*, volume 49. Elsevier, 2013.
[5] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
[6] Peixiang Zhong, Di Wang, and Chunyan Miao. An affect-rich neural conversational model with biased attention and weighted cross-entropy loss. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7492–7500, 2019.
[7] Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. Carer: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3687–3697, 2018.
[8] W John Wilbur and Karl Sirotkin. The automatic identification of stop words. *Journal of information science*, 18(1):45–55, 1992.
[9] Peter F Brown, Vincent J Della Pietra, Peter V Desouza, Jennifer C Lai, and Robert L Mercer. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–480, 1992.
[10] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.
[11] Saif M. Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL)*, Melbourne, Australia, 2018.
[12] Cristian Danescu-Niculescu-Mizil and Lillian Lee. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style

in dialogs. *arXiv preprint arXiv:1106.3077*, 2011.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[14] Sunghwan Mac Kim, Alessandro Valitutti, and Rafael A Calvo. Evaluation of unsupervised emotion models to textual affect recognition. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 62–70, 2010.

[15] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*, 2017.

[16] Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. Ambigqa: Answering ambiguous open-domain questions. *arXiv preprint arXiv:2004.10645*, 2020.

[17] Oana Bălan, Gabriela Moise, Livia Petrescu, Alin Moldoveanu, Marius Leordeanu, and Florica Moldoveanu. Emotion classification based on biophysical signals and machine learning techniques. *Symmetry*, 12(1):21, 2019.