Virtual Therapist
Using Deep Neural Network, Chatbot And Math Modalization in Mental Health Assessment

Nguyen Le Quoc Bao

Abstract

Adolescents are facing various challenges to their mental health due to their exposure to online risks and pressures, such as cyberbullying, misinformation, social comparison, and unrealistic expectations. They also have to deal with the stress and anxiety from their academic and personal lives, which can affect their self-esteem and well-being. These factors can lead to negative emotions and behaviors, such as depression, isolation, self-harm, and substance abuse. Teenagers are reluctant to ask for help because of the stigma around mental health disorders, and there are also additional limitations including waiting lists and geographic restrictions. There is a significant gap in the treatment that should be available conveniently and cost-effectively, and the services available at hand. The ratio of therapists, psychiatrists, psychiatric social workers and mental health nurses to patients is 1: 10,000, even in developed countries (Kislay, 2020). The disparity in the system means that most people with mental health problems will never get the support they need. In response, we developed a technology-based application that provides online support and guidance to adolescents with mental health problems.

1. Introduction

The application uses natural language processing and artificial intelligence to interact with users in a conversational manner, and offers a toolbox of features to help them cope with stress, anxiety, depression, and other challenges. The application also integrates mental health assessment tools to monitor the users' progress and provide feedback. We assume that technology-based applications can be a viable and scalable alternative to face-to-face mental health services for adolescents. Our solution consists of four main components:

- A general emotion classifier that can categorize the user's story (diary) into positive, negative, or neutral emotions, based on a deep neural network with bidirectional LSTM (BiLSTM) architecture.
- A complex emotion classifier that can further classify the user's story into 12 fine-grained emotion categories, such as anger, sadness, remorse, fear, depression, lonely, joy, love, optimism, gratitude, and pride. This component uses a transfer learning approach with BERT pretrained model to achieve better performance. The results from this step are then used to quantify the user's mental health quality based on a our mathematical formula.
- A chatbot that can respond to the user's story with empathetic and supportive messages, and suggest some practical solutions to help them cope with their negative emotions and improve their well-being.
- We also have a time series analysis model that can predict the user's future emotional trends based on their past diary entries. This component can help the user monitor their progress and identify potential risks or opportunities for intervention.

We evaluated our solution using a self-scraping dataset of online diaries from various websites. We compared different architectures and models for each component and selected the best ones based on their accuracy and performance.

2. Literary Reviews

2.1 Existing Dataset

Emotion classification is a challenging task that requires a large and diverse dataset to capture the nuances and complexity of human emotions. However, most existing datasets for emotion classification are limited in several ways. First, they only deal with a narrow range of emotions, often using only three categories: positive, negative, and neutral. This is not suitable for applications that need to differentiate between subtle and complex emotions, such as depression, anger, remorse, lonely, joy, pride, etc. For example, a teenager with deep depression should be more concerned than the one who just feels slightly sad. Second, most existing datasets use data from social media platforms, such as Twitter or Reddit (e.g GoEmotions - a dataset of fine-grained emotions), which are short and informal texts that may not reflect the true emotions of the users. Moreover, these texts may contain noise, such as slang, emojis, hashtags, or typos, that can affect the quality and reliability of the data. These datasets are not suitable for applications that require long and expressive texts, such as diary writing or storytelling. We want to build an application that allows adolescents to write their diary on each day and provide them with emotional feedback and support. Therefore, we need a dataset with the average length of sentence approximate to that of a diary entry, and that can capture the emotions based on the large context of story.

2.2 Existing Chatbots

Talking specifically about mental health chatbots, several implementations of mental health chatbots are available commercially, they include Wysa and Woebot. They are available in the form of android and IOS applications.

Wysa is an AI-based virtual therapist, who engages the user in a friendly dialogue using a blend of Cognitive behavioural therapy and mental health practices. Commercially available on Android phones and IOS systems as an application, Wysa protects the conversational data of the user by using encrypted chats and allows the user to use a concealed identity (van Aken, Betty, et al., 2019)

Another commercial implementation of a virtual therapist for mental
wellbeing is Woebot, which is also available in the form of an iPhone and Android application. It prompts the user to log in, unlike Wysa, and then initiates a short user survey where it tries to understand the user. It also practices Cognitive Behavioral Therapy, an approach to treatment which helps in improving mental state. The app provides regular check-ins for the user, short pre-filled options, and a gamified experience (van Aken, Betty, et al., 2019).

One of the limitations of the existing mental health chatbots is that they were launched before the advent of Chatgpt, a powerful natural language generation model that can produce coherent and fluent texts. Therefore, these chatbots cannot respond to the user in length, but rather seem like scripted and repetitive, which lower the user experience and engagement with the bot. We, instead, integrate Chatgpt to enhance our chatbot and make it more conversational and adaptive to the user's input. We also use the content in the chat to perform our task of emotion classification and mental health score calculation effectively. We believe that our chatbot can provide a better online support and guidance to adolescents with mental health problems.

---

3. Methodology

3.1 Applicatton Overview

Our website application allows you to post your status feeling and write your diary to talk to our chatbot, who can understand you without any judgment or secret revelation. Our chatbot can also help you improve your emotions if they are negative, or congratulate you if they are positive. Under the chat, we collect the conversations to perform analysis and produce a score of your mental health quality. This score reflects how well you are coping with your challenges and pressures, and how satisfied you are with your life. If this score continuously goes down and shows no sign of improving, the system will warn you by sending notifications, and suggest some resources and activities that can help you. If the situation is more serious, the system will send an emergency alert to trustworthy organizations that can intervene and act accordingly.

3.2 Data Scraping

To gather data for our website application, we used various sources and methods to collect texts or diaries of adolescents that reflect their emotions and mental health. We scraped data from several websites that allow users to share their personal stories and feelings.

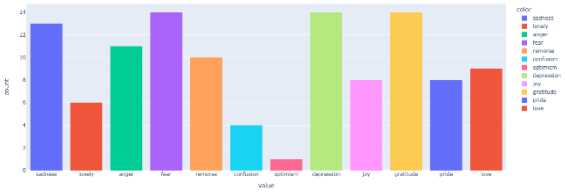

Figure 1: 12 classes of emotions distribution

Sentence	label
...I received a very low score on my math exam...	sadness
...to school is a nightmare for me because of demanding...	fear
... I just did something extremely tragically wrong...	remorse
...Sometimes the self-loathing gets the better of me...	depression
...in the dark of my room, seperated from the crowd...	lonely
...people continue to act in ways that I find repugnant...	anger

Table 1.1: Negative emotion collected from various website and conducted surveys

Sentence	label
...I have learned to appreciate the little things...	grateful
...I will win the first prize in the future...	optimism
...I arrived with flowers and said I'm here for you...	love
...Everyone is rooting for me and praising me...	pride
...same way that they could help others have pleasure...	joy
...It's kind of like a random rollercoaster of emotions...	confusion

Table 1.2: Postive & Neutral emotion collected from various website and conducted surveys

4. Deep Neural Network

4.1 Model 1

Due to the small-scale self-scraping dataset, we can try on various network and architecture to evaluate the model performace.

At first, with the task of classifying three general emotion categories (positive, negative, neurtral) we use RNN to capture the history of context for hidden state with SimpleRNN: $h_t = \gamma(W h_{t-1} + H z_t + b)$.

However, using only SimpleRNN does not yield a good performance. Instead we use LSTM (more robust with input gate, forget gate, memory cell) and combined with Bidirectional to capture the text in two directions $\hat{y}_t = \gamma\left(W^{(s)}\left[h_t^{(1)}, h_t^{(2)}\right] + c\right)$. The intial architecture with LSTM is rather light-weight and simple, but still acquire the best performance (converge after 20 epochs):

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_3 (Embedding)      (None, 200, 64)           640000

bidirectional_2 (Bidirectio  (None, 200, 128)          66048
nal)

bidirectional_3 (Bidirectio  (None, 64)               41216
nal)

dense_6 (Dense)              (None, 16)                1040

dense_7 (Dense)              (None, 3)                 51

=================================================================
Total params: 748,355
Trainable params: 748,355
Non-trainable params: 0
_________________________________________________________________
```

Figure 2.1 Model 1 architecture

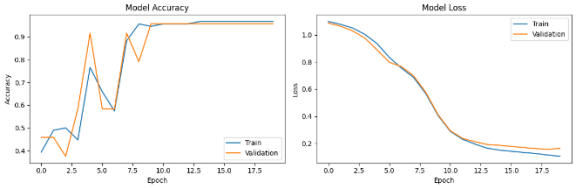

Figure 2.2: Training process through 20 epochs

Epochs	20
Accuracy on trainset	96.81%
Loss on trainset	0.1038
Accuracy on valset	95.83%
Loss on valset	0.167

Table 2.2: Final performance of model with BILSTM on 3 classes classfication.

4.2 Model 2

However, the above architecture does not perform well with the problem of 12-class classification. We need a more complex architecture with more layers to capture the slight nuances of emotions in texts. Accordingly, we choose to implement this architecture.

```
Layer (type)                 Output Shape              Param #
=================================================================
embedding_4 (Embedding)      (None, 200, 64)           640000

bidirectional_4 (Bidirectio  (None, 200, 128)          66048
nal)

bidirectional_5 (Bidirectio  (None, 200, 128)          98816
nal)

global_average_pooling1d (G  (None, 128)               0
lobalAveragePooling1D)

dense_8 (Dense)              (None, 500)               64500

dropout (Dropout)            (None, 500)               0

dense_9 (Dense)              (None, 200)               100200

dropout_1 (Dropout)          (None, 200)               0

dense_10 (Dense)             (None, 100)               20100

dropout_2 (Dropout)          (None, 100)               0

dense_11 (Dense)             (None, 50)                5050

dense_12 (Dense)             (None, 12)                612

=================================================================
Total params: 995,326
Trainable params: 995,326
Non-trainable params: 0
_________________________________________________________________
```

Figure 2.3: Model 2 architecture

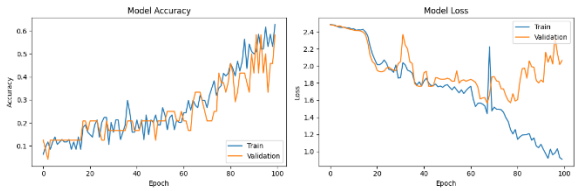

Figure 2.3: Training process through 100 epochs

Epochs	100
Accuracy on trainset	62.77%
Loss on trainset	0.9086
Accuracy on valset	58.33%
Loss on valset	2.0630

Table 2.3: Final performance of model with BILSTM on 12 classes classfication.

The performance of model is not good: accuracy on both train set and validation set fluctuates throughout 100 epochs. Furthermore, the loss on validation set starts increasing at epoch 80, showing the sign of overfiting. This means this model with above architecture (Figure 2.3) can not be continously trained.

4.3 Transfer Learning with BERT

Due to above obstacles, we decide to use a pretrained model and add to it with more layers with a hope that the large-scale pretrained BERT from Tensorflow can solve our problems.

```
Layer (type)                 Output Shape              Param #   Connected to
==================================================================================================
text (InputLayer)            [(None,)]                 0         []

preprocessing (KerasLayer)   {'input_word_ids': (None,  0         ['text[0][0]']
                              128),
                              'input_mask': (None, 128)
                             , 'input_type_ids': (None,
                              128)}

BERT_encoder (KerasLayer)    {'default': (None, 512),  4137318   ['preprocessing[0][0]',
                              'encoder_outputs': [(None  5         'preprocessing[0][1]',
                             , 128, 512),                         'preprocessing[0][2]']
                              (None, 128, 512),
                              (None, 128, 512),
                              (None, 128, 512),
                              (None, 128, 512),
                              (None, 128, 512),
                              (None, 128, 512),
                              (None, 128, 512)],
                              'sequence_output': (None,
                             128, 512),
                              'pooled_output': (None, 5
                             12)}

dense_16 (Dense)             (None, 400)               205200    ['BERT_encoder[0][9]']

dropout_13 (Dropout)         (None, 400)               0         ['dense_16[0][0]']

dense_17 (Dense)             (None, 200)               80200     ['dropout_13[0][0]']

dropout_14 (Dropout)         (None, 200)               0         ['dense_17[0][0]']

dense_18 (Dense)             (None, 100)               20100     ['dropout_14[0][0]']

dropout_15 (Dropout)         (None, 100)               0         ['dense_18[0][0]']

dense_19 (Dense)             (None, 50)                5050      ['dropout_15[0][0]']

dropout_16 (Dropout)         (None, 50)                0         ['dense_19[0][0]']

dense_20 (Dense)             (None, 30)                1530      ['dropout_16[0][0]']

dropout_17 (Dropout)         (None, 30)                0         ['dense_20[0][0]']

classifier (Dense)           (None, 12)                372       ['dropout_17[0][0]']

==================================================================================================
Total params: 41685637 (159.02 MB)
Trainable params: 312452 (1.19 MB)
Non-trainable params: 41373185 (157.83 MB)
```

Figure 2.4: Model 2 architecture with pretrained BERT from Tensorflow

The version of BERT for “preprocessing_layer” is “bert_en_uncased_preprocess/3” and that of “encoder” is “bert_en_uncased_L-8_H-512_A-8/2”.

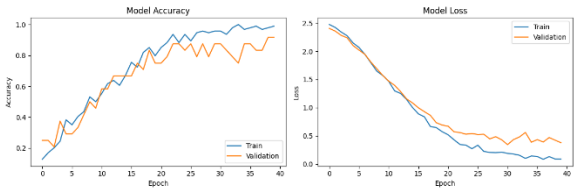

Figure 2.4: Training process through 40 epochs

Epochs	40
Accuracy on trainset	98.94%
Loss on trainset	0.0897
Accuracy on valset	91.67%
Loss on valset	0.3791

Table 2.4: Final performance of model with pretrained BERT on 12 classes classfication.

We stop early at epoch 40 when validation loss stops declining to prevent overfitting. However, at epoch 40, the accuracy on train and validation set is over 90%, which is really good.

5. Mental Health Quality Assessment

(This math formula is devised by Nguyen Le Quoc Bao)

We define a math formula to assess the mental health quality assessment for users based on some most trustworthy and popular mental health quality scale on the world such as DASS.

In a very simple idea, an adolescent with depression should be more concerned than the one who just only feel lonely or sadness. Therefore, we assign each category with a score number indicating how serious such emotion is. Then we calculate the temporary score is the mean of 3 classes with greatest probabilities. Remember that an input text is just one chat from user, but our user will continue to chat with the bot and maybe, at the end of chat, things will change: user may get more relieved or happy thanks to the help from chatbot. Due to such reason, it's not sensible to regard the mean score of all chat as the official score for user's mental health quality. Hence, we use $\gamma \in (0,1)$ as a regulator: the more recent the score, the more impactful it will be.

$$\hat{y}^i = \text{classifier\_model.predict}([\text{chat}^i]) \mid \text{with } i = 1...N$$

$$\hat{y}_1^i, \hat{y}_2^i, \hat{y}_3^i = \arg\max^3(y^i)$$

$$s_i = \frac{1}{3}\sum_{j=1}^{3} f(classes[\hat{y}_j^i])$$

$$Q = \gamma^{N-1} * s_i + \gamma^{N-2} * s_{i+1} + \ldots + \gamma^0 s_N$$

$$Q = \sum_{i=1}^{N} \gamma^{N-i} * s_i$$

Explain my formula: $\hat{y}^i$ is the probability vector returned from the "classifier_model" with the input is the i-th chat of total N chats from user in the conversation. Next, we take 3 greatest probabilties instead of just one. When we have the three indices, we can get the score of each with the mapping function f(e.g depression gets $5pts$, anger gets $4pts$, sadness gets $3pts$...) and then we get the mean of three → we get s_i (score of i-th chat). Now after user complete all N chats with the bot, we calculate the final score value Q (quality). γ^{N-1}means the value s_i will be decreased, while the final value s_N still remains the same.

6. Future Work & Conclusions

In this paper, we have presented a technology-based solution that can provide online support and guidance to adolescents with mental health problems. Our solution consists of four main components: a general emotion classifier, a complex emotion classifier, a mental health score calculator, and a chatbot. We have evaluated our solution using a self-scraping dataset of online diaries, and conducted a user study to assess its usability and effectiveness. However, our solution also has some limitations and challenges that we plan to address in our future work. Some of the directions for future work are:

- We will collect more data from various sources and domains to make our dataset more diverse and representative of the emotions and mental health issues of adolescents. We will also use active learning and data augmentation techniques to increase the size and quality of our dataset.
- We will expand the emotion categories to reflect more nuanced and complex moods, such as frustration, boredom, curiosity, or excitement. We will also explore the use of dimensional models of emotions, such as valence, arousal, and dominance, to capture the intensity and variability of emotions.
- We will update our models and web application with the latest advancements in natural language processing and artificial intelligence, such as transformers, pre-trained language models, and multimodal learning. We will also optimize our models and web application for speed, scalability, and security.
- We will conduct more user studies with different groups of adolescents, such as those from different cultures, backgrounds, or contexts, to evaluate the generalizability and adaptability of our solution. We will also collect user feedback and suggestions to improve the design and functionality of our solution.

We hope that our future work will further enhance our solution and make it more useful and accessible for adolescents who need online support and guidance for their mental health and well-being.