Journal of Intelligent Information Systems
https://doi.org/10.1007/s10844-022-00704-w

Multimodal time-aware attention networks for depression detection

Ju Chun Cheng[1] · **Arbee L. P. Chen**[2]

Received: 4 October 2021 / Revised: 17 March 2022 / Accepted: 20 March 2022

Abstract
Depression is a common mental disorder, which may lead to suicide when the condition is severe. With the advancement of technology, there are billions of people who share their thoughts and feelings on social media at any time and from any location. Social media data has therefore become a valuable resource to study and detect the depression of the user. In our work, we use Instagram as the platform to study depression detection. We use hashtags to find users and label them as depressive or non-depressive according to their self-statement. Text, image, and posting time are used jointly to detect depression. Furthermore, the time interval between posts is important information when studying medical-related data. In this paper, we use time-aware LSTM to handle the irregularity of time intervals in social media data and use an attention mechanism to pay more attention to the posts that are important for detecting depression. Experiment results show that our model outperforms previous work with an F1-score of 95.6%. In addition to the good performance on Instagram, our model also outperforms state-of-the-art methods in detecting depression on Twitter with an F1-score of 90.8%. This indicates the potential of our model to be a reference for psychiatrists to assess the patient; or for users to know more about their mental health condition.

Keywords Depression detection · Social media · Deep learning

1 Introduction

Depression is a common mental disorder, which affects 264 million people worldwide (James et al., 2018). Symptoms of depression include feeling sad most of the time and los-

✉ Ju Chun Cheng
cjuchun@gmail.com

Arbee L. P. Chen
arbee@asia.edu.tw

[1] Institute of Information Systems and Applications, National Tsing Hua University, Hsinchu, Taiwan

[2] Department of Computer Science and Information Engineering, Asia University/Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan

Published online: 18 April 2022

ing interest in usual activities, which not only influence one's performance at work or school but also impact the relationships with family and friends. When the condition is severe, depression sufferers may even have the thought or behavior of self-harm or suicide. Therefore, being correctly diagnosed and given sufficient treatment is important to prevent the condition from getting worse. However, only a few of the depression sufferers are treated (Wang et al., 2007), the reasons include lack of resources, unaware of the illness, and having the social stigma of mental illness. Traditionally, screen tools (e.g., CES-D (Radloff, 1977), PHQ-9 Kroenke et al., 2001, and BDI-II Beck et al., 1996) and face-to-face interviews with psychiatrists are needed to diagnose depression, which is expensive and time-consuming. To tackle this problem, De Choudhury et al. (2013) examined the possibility of using social media as a tool to predict and detect depression.

With the advancement of technology, billions of people are using social media including Twitter, Facebook, Instagram, etc. Through social media, people can share their thoughts and feelings anytime and anywhere. Therefore, we can learn one's mental state through the use of languages, posted images, and online behaviors on social media. Conventionally, diagnosis of depression depends on patients' self-report of feelings in the past few weeks, which is limited in granularity since what they report is a summary of feelings over a long period of time. Therefore, social media is a good platform to provide extra and finer-grained information to study depression.

Earlier studies mainly used the information from texts to detect depression (Coppersmith et al., 2014; De Choudhury et al., 2013). Reece and Danforth (2017) found that images posted on social media can also be used to distinguish depressed and non-depressed users. Moreover, we can better understand how a user feels by considering textual and visual information together (Gui et al., 2019). Besides texts and images, posting behaviors are also discriminative for detecting depression (Shen et al. 2017).

Although many features were explored in the studies of depression detection, only a few works have considered the information of the time gaps between posts. Time intervals are important information when it comes to health-related analysis. Take electronic health record data as an example, time intervals between visits or admissions to the hospital vary from years to days. Frequent visits or admissions may indicate a severe health condition. Similarly, when the time gaps between depressed moods are long it means the user feels sad once in a while. On the other hand, if the user feels depressed frequently it may indicate there exists some mental health problems.

Many works detected depression effectively using deep learning models. However, most of them did not provide the interpretability of their models. Without interpretability, it may be difficult for the users, e.g., psychiatrists, psychologists, or social media users, to trust or further use the results from the models, especially in the healthcare domain.

To tackle the above challenges, we proposed a model named *multimodal time-aware attention networks* (MTAN), which not only utilizes the information from texts, images, and posting time but also considers the time intervals between posts. Furthermore, we used an attention mechanism to pay more attention to the posts that are important for depression detection and provide interpretable results.

The process of our work is briefly stated as follows. We used hashtags to find users on Instagram and labeled them according to their self-statement. After scraping the posts, we extracted and concatenated text, image, and posting time features of each post. The post representations are then passed through the time-aware LSTM. In the time-aware LSTM unit, the previous memory is adjusted by the elapsed time between posts. After that, we used an attention layer to let the model decide to pay more or less attention to each post

and obtain the user representation. Finally, a fully connected layer with a sigmoid activation function is used to get the prediction result. Our contributions in this paper are as follows:

- We incorporate text, image, and posting time features of posts and propose a multimodal depression detection model which uses time-aware LSTM to consider time interval information and an attention mechanism to pay more attention to the relevant posts. Experiment results demonstrate that our model reaches an F1-score of 95.6%.
- Our model can also obtain a promising result in detecting depression on Twitter, which outperforms other works with an F1-score of 90.8%.
- In addition to detecting depression, our model also provides interpretable results by visualizing the attention weight of the posts and give insight into which posts contribute to the model prediction.

The rest of the paper is organized as follows: related works are reviewed in Section 2, the task is formulated in Section 3, details of MTAN are explained in Section 4, results of the experiment are presented in Section 5, and the conclusion and future work of the paper are provided in Section 6.

2 Related work

After Park et al. (2012) demonstrated the possibility of using social media data to study depression, many works examined the potential of using social media to detect depression and get promising results.

Different social media platforms are used to study depression detection. Many works used the data from Twitter (De Choudhury et al., 2013; Gui et al., 2019), some other works studied depression detection on Facebook (Wu et al., 2020), Reddit (Yates et al., 2017), and Weibo (Shen et al., 2018). In our work, we used Instagram, the most used social media among the age of 18-44 in Taiwan[1], as the platform for depression detection. Among the works that used Instagram for depression detection, we have collected the most users which is at least twice more than the other works.

There are different ways to collect and label data from social media for depression detection. De Choudhury et al. (2013) employed crowdsourcing to collect Twitter users and took the results of their CES-D test as ground truth. Although this method can obtain reliable data, the number of collected users is limited. Hence, Coppersmith et al. (2014) proposed to use regular expressions to find Twitter users who had self-stated being diagnosed with depression and proved the effectiveness of their data collecting approach by demonstrating that four different mental illnesses can be classified by the statistical model. This data collecting method not only requires less cost and effort but also can acquire more samples.

Earlier studies (De Choudhury et al., 2013; Coppersmith et al., 2014) used hand-crafted features such as engagement, emotion, linguistic style, demographic, ego-network, etc., which mostly are from text and behavior of the users, and used statistical models to detect depression. Reece and Danforth (2017) extracted features from images, e.g., number of faces in the photo, HSV (Hue, Saturation, and Value), and type of filters, and used Random Forests classifier as their predictive model. These studies demonstrated that both textual

[1] https://napoleoncat.com/stats/social-media-users-in-taiwan/2021/06

and visual features are effective to detect depression. Recently, researchers started to utilize multimodal features with deep learning methods for depression detection. Gui et al. (2019) incorporated textual and visual information and proposed a reinforcement learning method to select texts and images that are indicative of depression. The experiment results demonstrated that using multimodal information is better than using a single modality. In addition to texts and images, posting behaviors also convey important information. Shen et al. (2017) examined the effectiveness of different modalities and found that the performance drops the most when removing the posting behaviors. In our work, we used texts, images, and posting time jointly to detect depression.

While a range of features was explored to detect depression, most of the works did not consider the information of time intervals between posts. Based on the depression criterion defined by Diagnostic and Statistical Manual of Mental Disorders (DSM) (American Psychiatric Association, 2013), one must experience depressed moods most of the day, nearly every day for at least two weeks to be diagnosed as depression. It means that the time interval between depressed moods of depression sufferers should be shorter than the non-depressed ones who feel sad occasionally. Chiu et al. (2021) first considered time intervals in detecting depression with social media data. First, the depression scores of the posts from a user are predicted as the origin depression scores. Then, the origin depression score of the t^{th} post is aggregated as the weighted sum of the $t-1^{th}$ aggregated depression score and the t^{th} origin depression score, in which the weight is adjusted according to the time interval between posts. Afterward, if the average score in a day is over a threshold, the day will be identified as a depressive day and the users will be detected as depressive if half of the posting days are depressive. Different from this method, we used a modified LSTM which takes time intervals between posts into consideration to detect depression. LSTM (Hochreiter & Schmidhuber, 1997) is an RNN variant, which can capture long-term dependencies in sequential data. However, the standard LSTM cannot handle time irregularities in sequences. Baytas et al. (2017) modified the LSTM architecture, which takes time intervals into consideration to adjust the previous memory.

Although recent studies are making great progress using deep learning methods, only a few of them provide interpretable results of their models. The attention mechanism is often used to get the interpretability of the model, which was initially utilized in machine translation (Bahdanau et al., 2015). Vaswani et al. (2017) used self-attention to relate different positions of a single sequence to compute the representation of a sequence.

Inspired by the aforementioned works, we proposed to use multimodal data and employ time-aware LSTM to take time intervals information into account. After getting the time-adjusted hidden states of each post, we then exploit an attention mechanism to find the relevant posts for detecting depression.

3 Problem formulation

In social media data, each user $u \in U$ can be represented as a sequence of posts $P = \{p_1, p_2, ..., p_n\}$, where n is the total number of posts and p_t is the t^{th} post. We detect depression using multimodal features, i.e., texts, images, posting time, and the time interval between the posts, that is, $p_t = (txt_t, img_t, time_t, \Delta t_t)$. Given a sequence of posts of a user u_i, our goal is to train a model that can predict the label y_i for the user, where $y_i \in \{1, 0\}$.

4 Method

In this section, we first introduce the process of data collection. Then we describe how we preprocess the collected data. Next, we explain the procedure of feature extraction. Finally, we illustrate the detail of depression detection.

4.1 Data collection

To train the model for the task of depression detection on social media, we constructed a dataset with depressive and non-depressive users via Instagram. The data collection process is illustrated in Fig. 1.

Different from Twitter, in which you can search for the tweets containing specific words or sentences, one can only search for usernames, hashtags, and places on Instagram. Therefore, the data collection method for Twitter cannot be directly applied to Instagram.

To collect depressive users on Instagram, we used depression-related words (e.g., depressed (憂鬱、抑鬱), depression (憂鬱症、抑鬱症、重鬱症)) as hashtags to find posts that contain these hashtags. For the users of the found posts, we retrieve their posts within one year of the latest posting date. Next, inspired by Coppersmith et al. (2014), users are labeled as depressive if they have self-stated being diagnosed with depression in their biography or posts. Furthermore, we found that many depressive users follow each other on Instagram. Therefore, we scraped the followers and followees of the depressive users and retrieve their posts and label them with the rules mentioned above. Finally, we obtained 526 depressive users and 20,618 posts.

For non-depressive users, Chiu et al. (2021) and Huang et al. (2019) used "#happy" to find non-depressive users, which may find users who are more optimistic and post more positive posts. Instead, we used "#daily (日常)" and "#life (生活)" to find users who tend to post about their daily life. For the users of the posts we found with hashtags, we scraped their posts within one year of the latest posting date. We then checked their biography and posts to remove those who have self-identified being diagnosed with depression.

For each post, we scraped the caption, images, and posting time of the post. In Instagram, each post must have at least one image and can have up to 10 images. Different from previous works of detecting depression on Instagram (Chiu et al., 2021; Huang et al., 2019; Mann et al., 2020), which only took one image for each post, we scraped all the images in a post.

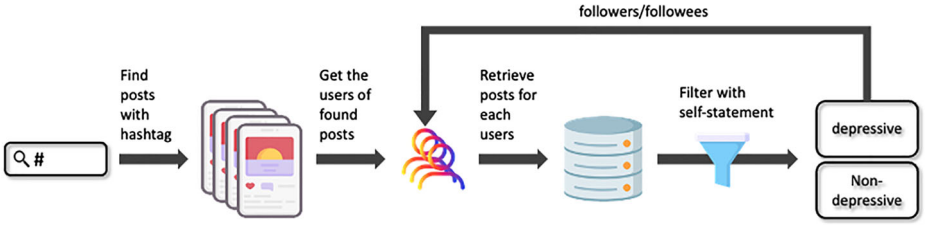

Fig. 1 Flow chart of data collection

4.2 Data preprocessing

In order to have better performance, we preprocessed texts and images of each post before feature extraction.

4.2.1 Text preprocessing

Since hyperlinks and @mentions do not convey much information, we removed them from the texts. For the reason that we used hashtags to find depressive and non-depressive users, we removed all the hashtags in the posts so that hashtags information is not taken into account when classifying the users. According to Novak et al. (2015), tweets with and without emojis have different sentiment meanings. Instead of removing emojis, we replaced emojis with their corresponding meanings.

4.2.2 Image preprocessing

Based on the input requirement of InceptionResNetV2, we resized each image to 299x299 and normalized the pixel values of each image by rescaling them between -1 and 1.

4.3 Feature extraction

After texts and images are preprocessed, we extracted text features and image features automatically using pre-trained models. In contrast, posting time features are extracted as predefined features, i.e., seasons, days of a week, and parts of a day.

4.3.1 Text feature extraction

BERT (Devlin et al., 2019) is a transformer-based model for natural language processing, which was pre-trained on a large corpus and can be fine-tuned to achieve state-of-the-art performance in many NLP tasks, e.g., question answering and natural language inference. The input of BERT is a sequence of tokens, in which the first token is always the special symbol "[CLS]." The final hidden state of [CLS] is the aggregated representation of the sequence.

Given the t^{th} text txt_t of each user, we encoded the text with pre-trained BERT and obtained the hidden representation of [CLS], $\hat{h}_t \in R^{768}$. The text feature representation of the t^{th} post $h_t^{txt} \in R^d$ is then computed by passing the hidden representation of [CLS], $\hat{h}_t$, through a fully connected layer followed by a hyperbolic tangent (tanh) activation function:

$$h_t^{txt} = tanh(W_{txt}\hat{h}_t + b_{txt}).$$

4.3.2 Image feature extraction

InceptionResNetV2 (Szegedy et al., 2017) is a convolutional neural network (CNN) that combines the Inception architecture with residual connection, which achieved a top-5 accuracy of 95.3% on the ImageNet (Deng et al., 2009) validation dataset.

We used pre-trained InceptionResNetV2 to extract image features. Given the image img_t of the t^{th} post, we took the output vector $\hat{h}_t \in R^{1536}$ from the last average pooling layer in InceptionResNetV2 to calculate the image feature representation $h_t^{img} \in R^d$ as $h_t^{img} = tanh(W_{img}\hat{h}_t + b_{img})$.

For the post that have multiple images $img_t^1, img_t^2, ..., img_t^n$, where n is the number of images in the t^{th} post and img_t^i is the i^{th} image in the t^{th} post, we used pre-trained InceptionResNetV2 to extract image vectors $\hat{h_t^1}, \hat{h_t^2}, ..., \hat{h_t^n}$. Then we passed the average image vectors $\hat{h}_t = (\hat{h_t^1} + \hat{h_t^2} + ... + \hat{h_t^n})/n$ through a fully connected layer followed by a tanh activation function to compute the image feature representation of the t^{th} post h_t^{img}:

$$h_t^{img} = tanh(W_{img}\hat{h}_t + b_{img}).$$

4.3.3 Posting time feature extraction

After analyzing user posting time, we found that the distribution of posting season, posting day, and posting hour are distinctive between depressive users and non-depressive users. Details of the data analysis are presented in Section 5.2.

For the posting time feature representations, the season, the day of a week, and the part of a day are extracted from the posting time for each post. The 4 seasons, 7 days of a week, and 4 parts of a day are first one-hot encoded and then passed through fully connected layers respectively to get the representations h_t^{sea}, h_t^{day},and h_t^{part}. The details of the posting time features are presented in Table 1.

4.3.4 Modality fusion

To incorporate textual, visual, and time information for depression detection, we concatenated the text feature representation, image feature representation, and posting time feature representations of the t^{th} post to get the post representation X_t of the t^{th} post:

$$X_t = h_t^{txt} \oplus h_t^{img} \oplus h_t^{sea} \oplus h_t^{day} \oplus h_t^{part}.$$

Table 1 Posting time features

Posting Time Features	
Season	Spring (March - May)
	Summer (June - August)
	Fall (September - November)
	Winter (December - February)
Day of a week	Monday
	Tuesday
	Wednesday
	Thursday
	Friday
	Saturday
	Sunday
Part of a day	Midnight (0:00 - 5:59)
	Morning (6:00 - 11:59)
	Afternoon (12:00 - 17:59)
	Evening (18:00 - 23:59)

4.3.5 Time interval

Given the t^{th} post, the time interval of the t^{th} post Δ_t is calculated by subtracting the posting time of the $t-1^{th}$ post PT_{t-1} from the posting time of the t^{th} post PT_t:

$$\Delta_t = PT_t - PT_{t-1},$$

where Δ_t is measured in days, and $\Delta_1 = 0$.

4.4 Depression detection

After the text features, image features, and posting time features of a post were concatenated to obtain the post representation, all the post representations of a user were fed into the time-aware LSTM layer and the attention layer to get the user representation. Finally, we used a fully connected layer with a sigmoid activation function to get the possibility of depression. The architecture of MTAN is shown in Fig. 2.

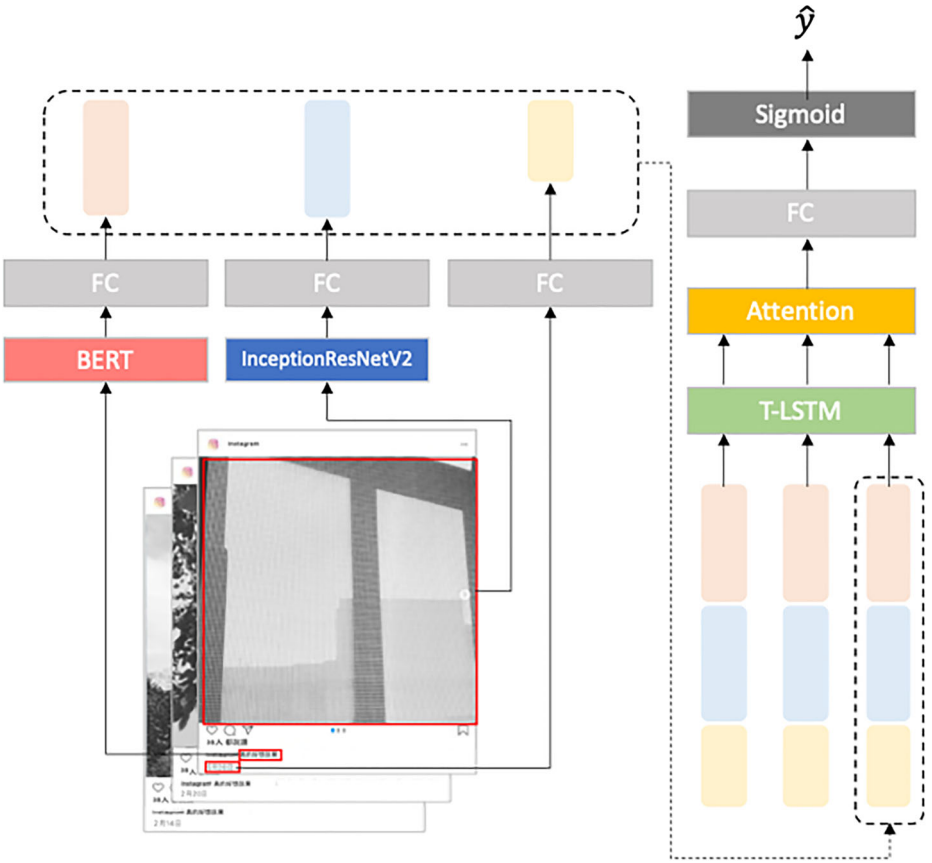

Fig. 2 Architecture of MTAN

4.4.1 Time-aware LSTM

RNN is a powerful method for modeling sequential data like social media data. However, RNN has the problem of vanishing and exploding gradients. LSTM (Hochreiter & Schmidhuber, 1997) is a variant of RNN that can tackle the aforementioned problem and handle the long-term dependencies of the sequential data via a gated architecture. An LSTM unit is composed of an input gate, a forget gate, an output gate, and a memory cell. The current cell memory C_t is obtained by forgetting the previous cell memory C_{t-1} through a forget gate f_t and adding new information from the candidate memory $\tilde{C}_t$ through an input gate i_t. Finally, the current hidden state h_t is generated by filtering current cell memory through an output gate o_t:

$$
\begin{aligned}
f_t &= sigma;\,(W_f x_t + U_f h_{t-1} + b_f) && \text{(Forget gate)}\\
i_t &= sigma;\,(W_i x_t + U_i h_{t-1} + b_i) && \text{(Input gate)}\\
o_t &= sigma;\,(W_o x_t + U_o h_{t-1} + b_o) && \text{(Output gate)}\\
\tilde{C}_t &= tanh(W_c x_t + U_c h_{t-1} + b_c) && \text{(Candidate memory)}\\
C_t &= f_t * C_{t-1} + i_t * \tilde{C}_t && \text{(Current memory)}\\
h_t &= o_t * tanh(C_t), && \text{(Current hidden state)}
\end{aligned}
$$

where x_t is the current input, h_{t-1} is previous hidden state, and {W,U,b} are the network parameters to be trained.

However, the standard LSTM cannot handle the input sequence that has irregular time intervals. The architecture of LSTM implicitly assumes the inputs are distributed uniformly, which is not the characteristic of social media data. Time gaps between social media posts vary from seconds to years. Different time intervals between posts may contain different information. Therefore, we use time-aware LSTM which incorporates time interval information into the standard LSTM architecture. The main difference of time-aware LSTM is that the previous cell memory is adjusted proportionally by the elapsed time. Since we do not want to lose the overall profile of the user, only the short-term memory is discounted while the long-term memory remains the same. A non-increasing function is used to transforms the elapsed time into the weight to discount the short-term memory.

In the time-aware LSTM unit, the previous memory C_{t-1} is decomposed into the short-term memory C^S_{t-1} and the long-term memory C^L_{t-1} with a network. The short-term memory is then discounted by the elapsed time through a non-increasing function. Next, the long-term memory and the discounted short-term memory $\hat{C}^S_{t-1}$ are combined to compose the adjusted previous memory C^*_{t-1}. The current memory C_t is then calculated using the adjusted previous memory C^*_{t-1} instead of the previous memory C_{t-1}. The architecture of the time-aware LSTM unit is illustrated in Fig. 3. Detailed mathematical expressions are as below:

$$
\begin{aligned}
C^S_{t-1} &= tanh(W_d C_{t-1} + b_d) && \text{(Short-term memory)}\\
\hat{C}^S_{t-1} &= C^S_{t-1} * g\Delta_t && \text{(Discounted short-term memory)}\\
C^L_{t-1} &= C_{t-1} - C^S_{t-1} && \text{(Long-term memory)}\\
C^*_{t-1} &= C^L_{t-1} + \hat{C}^S_{t-1}, && \text{(Adjusted previous memory)}
\end{aligned}
$$

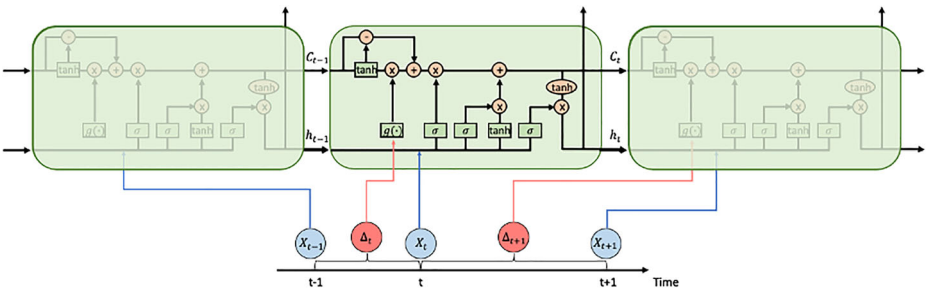

Fig. 3 Time-aware LSTM unit (Baytas et al., 2017)

where C_{t-1} is the previous cell memory, W_d and b_d are the parameters of subspace decomposition, Δ_t is the time interval between posts, and $g()$ is a heuristic decaying function that reduces the effect of short-term memory as Δ_t increases. We choose $g(\Delta_t) = 1/(log(e + \Delta_t))$ as Baytas et al. (2017) suggested.

4.4.2 Attention mechanism

Different posts contribute differently to the decision of depression detection. Therefore, we use an attention mechanism to extract the posts that are important for detecting depression.

We used self-attention as our attention mechanism, in which the hidden state of the t^{th} post is the query, key, and value of the t^{th} post. The dot products of queries and keys are first computed and then scaled and normalized to get the attention weights. The weights are then used to create a linear combination of values to get the output of each post. The outputs are then gone through a global average pooling layer to obtain the user representation v.

4.4.3 Depression detection

After getting the user representation v from the attention layer, it is then passed through a fully connected layer followed by a sigmoid activation function to get the probability of depression:

$$\hat{y} = sigmoid(W_u v + b_u).$$

5 Experiments

In this section, we first describe the dataset and analyze the statistics of the dataset. Then we introduce the setup of our experiments. Next, we compare different feature extraction models. After that, we present the results of different experiments. Finally, we prove the effectiveness of MTAN with real cases.

5.1 Dataset

With the data collection methods described in Section 4.1, 1054 users were collected from Instagram, which includes 526 depressive users and 528 non-depressive users. Table 2 illustrates the statistics of the dataset.

Table 2 Statistical detail of the dataset

	Users	Posts
depressive	526	20618
non-depressive	528	23772

5.2 Data analysis

In this section, we analyze the posting behavior of depressive and non-depressive users. We compare the difference of posting time, posting day, and posting season.

5.2.1 Posting time

Figure 4 illustrates the posting time distribution of depressive users and non-depressive users. The trend of the posting time is similar between the two groups. However, the posting ratios of the two groups are significantly different from 9pm to 6am (p-value$<$0.05). The result complies with the depression criteria defined by DSM that depression sufferers have trouble falling or staying asleep so they are more active in the middle of the night.

5.2.2 Posting day

Figure 5 shows the posting day distribution of depressive users and non-depressive users. Non-depressive users tend to post more posts on weekends. In contrast, the distribution of depressive users does not fluctuate much during the week. Most of the people work or go to school on weekdays and go out on weekends. They tend to post about what they do, where they go, and what they eat when they go out. However, depression sufferers have lost interest and pleasure in doing anything so they tend to stay home during the week and post about what they feel and their depression experiences.

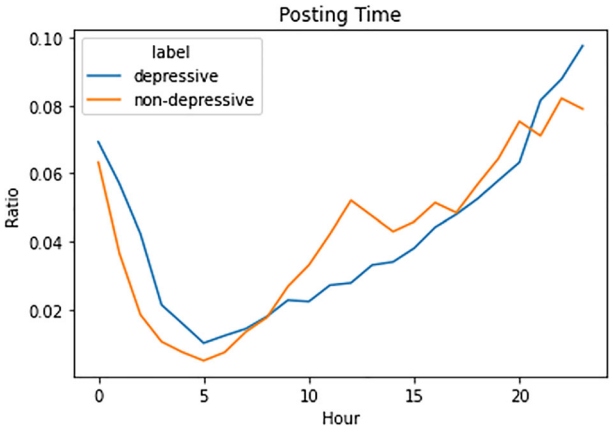

Fig. 4 Posting time distribution

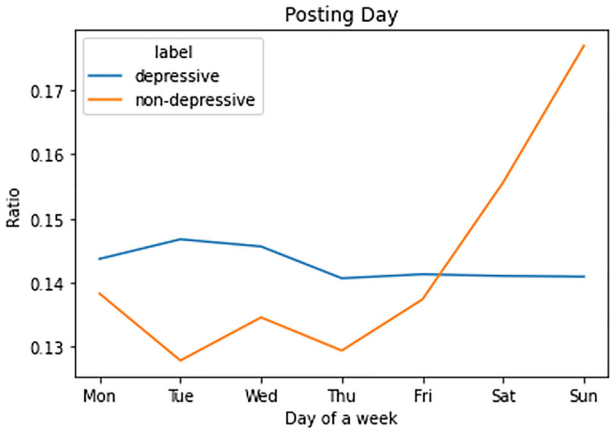

Fig. 5 Posting day distribution

5.2.3 Posting season

Figure 6 is the pie chart of the posting season of the two groups. In fall and winter, depressive users have more posts than non-depressive users. Some depression sufferers are more likely to become depressed in fall and winter because of the lack of sunlight.

5.3 Experiment setup

We used grid search to explore the hyperparameters: the dimension of the time-aware LSTM hidden state {64, 128, 256, 512}, the learning rate {0.001, 0.0001} and the batch size {32, 64, 128}. We conducted 5-fold cross validation experiments on the explored hyperparameters. The optimal dimension of the time-aware LSTM hidden state, batch size and learning rate are 256, 64 and 0.001, respectively. Adam (Kingma & Ba, 2015) optimizer was used to optimize the model and the binary cross-entropy was used as the loss function. We implemented the methods with Tensorflow2 (Abadi et al., 2016) and trained the models on NVIDIA Tesla V100.

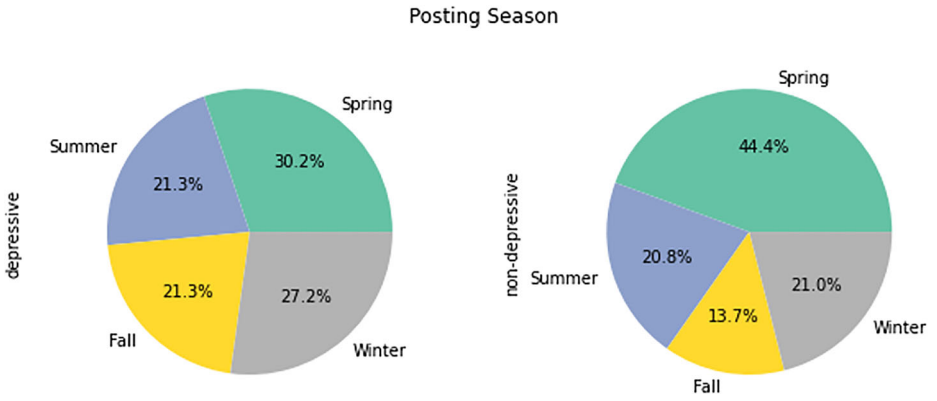

Fig. 6 Posting season

Table 3 Performance of different word embedding method

Word Embedding	F1
Word2Vec	0.902
fastText	0.897
GloVe	0.906
BERT	**0.912**

5.4 Feature extraction model comparison

We have performed different experiments on extracting text, image, and posting time features. In this section, we demonstrate the results of using different methods to extract text, image, and posting time features.

5.4.1 Word embedding

In our work, we used BERT to obtain the word embedding from the social media text. We compared the performance of BERT with three commonly used word embedding methods, which are Word2Vec, fastText, and GloVe. Before getting the embedding of the text, we preprocessed the text as described in Section 4.2, and then segmented the text using CKIP[2]. After getting the embedding of the segmented words, the word embeddings were then going through a bidirectional LSTM to get the post embeddings. After that, we passed the post embeddings of a user through the time-aware LSTM and attention layer to get the user embedding. Finally, we used a fully connected layer with a sigmoid activation function to get the prediction result. We achieved around 90% F1-score with Word2Vec, fastText, and GloVe. However, the F1-score can get at least 0.6% higher when using BERT (Table 3).

5.4.2 CNN model

In our method, we used InceptionResNetV2 to extract image features from social media images. We compared the performance of InceptionResNetV2 with VGG16 and ResNet50, which are the two CNN models that many other works used. The CNN models we used were pre-trained on the ImageNet dataset. Before using VGG16 and ResNet50 to extract image features, we resized the images to 224x224 and converted the images from RGB to BGR then subtracted ImageNet average BGR. The performance of the three CNN models was similar, all of them achieved around 90% F1-score. However, we chose to use the one that can get the highest F1-score (Table 4).

5.4.3 Posting time features

As described in Section 4.3.3, we extracted the features from posting time and obtained the posting time features for each post. Instead of getting the posting time features on post-level, we tried to use the posting time features on user-level to compare the performance. The user-level posting time features were obtained by calculating the proportions of posts posted in 4 seasons, 7 days of a week, and 4 parts of a day of the entire posting history of a user. As illustrated in Fig. 7, the user-level posting time features are concatenated with the user

[2]https://github.com/ckiplab/ckip-transformers

Table 4 Performance of different CNN model

Word Embedding	F1
VGG16	0.891
ResNet50	0.898
InceptionResNetV2	**0.901**

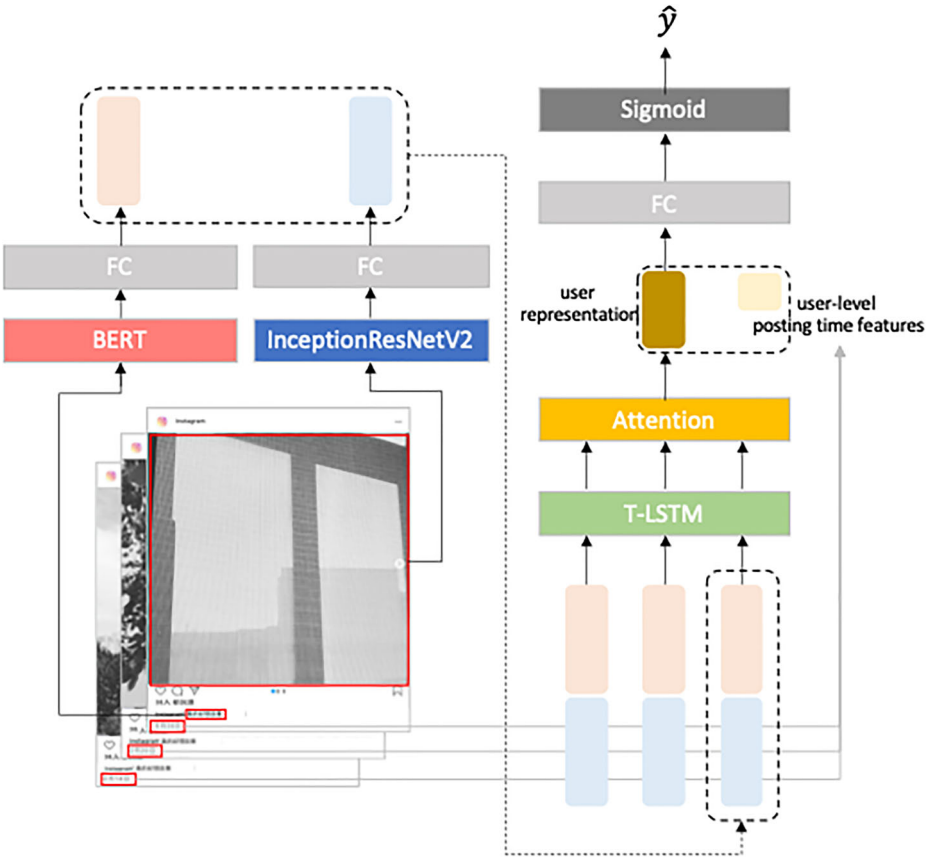

Fig. 7 Architecture of using user-level posting time features

Table 5 Performance of using different level of posting time features

	F1
User-level posting time features	0.939
Post-level posting time features	**0.956**

Table 6 Comparison of mechanism effectiveness

	Precision	Recall	F1
Chiu et al.	0.899	0.856	0.877
MLN	0.934	0.907	0.920
MLN + attentiont	0.938	0.935	0.936
MLN + time-aware	0.943	0.943	0.943
MTAN	**0.950**	**0.963**	**0.956**

representation and goes through a fully connected layer with a sigmoid activation function to get the prediction result. However, the performance was better when using the post-level posting time features. By using post-level posting time features, the emotion of each post can be better understood so that a user can be correctly detected as depressive (Table 5).

5.5 Results and analysis

In this section, we first compare the performance with other work. Then we run an ablation study on the proposed mechanisms. Next, we evaluate the effectiveness of different modalities. After that, we discuss the results of using different data collection periods. Finally, we test our method on a Twitter multimodal dataset.

5.5.1 Method comparison

We compared MTAN with the method proposed by Chiu et al. (2021). We implemented the method and trained the model on our data. Chiu et al. first predicted the depression score of each post then used a two-stage detection mechanism that aggregates the score by time-adapted weight and applies day-based detection to detect depressive users. Instead of detecting depression in three stages, we used time-aware LSTM followed by an attention mechanism to detect depression in one stage. MTAN outperforms Chiu et al. by 7.9% in F1-score.

We used *multimodal LSTM networks* (MLN) as the base model and added a time-aware mechanism and an attention mechanism respectively to compare their effectiveness. As we can see from the experiment results, the attention mechanism and the time-aware mechanism are both effective and can improve the performance by 1.6% and 2.3% in F1-score.

Table 7 Comparison of different modality combination

	Precision	Recall	F1
text	0.925	0.932	0.928
image	0.875	0.860	0.867
posting time	0.686	0.713	0.699
text + image	0.949	0.940	0.944
text + posting time	0.936	0.933	0.934
image + posting time	0.878	0.886	0.882
text + image + posting time	**0.950**	**0.963**	**0.956**

Table 8 Comparison of using single or multiple images

	Precision	Recall	F1
single image	0.944	0.958	0.950
multi-image min	0.957	0.934	0.945
multi-image max	0.946	0.942	0.943
multi-image avg	**0.950**	**0.963**	**0.956**

Using the time-aware LSTM to adjust the memory of the previous post by the time interval of the posts is more effective than using an attention mechanism to give the *depressive* posts, i.e., the posts containing texts or images with depressive moods higher attention weights. Overall, utilizing the time-aware mechanism and the attention mechanism jointly gets the best result. The result of the recall is higher indicating most of the depressive users can be detected by our model (Table 6).

5.5.2 Modality effectiveness

We evaluated the contribution of different modalities by comparing the results of different modalities combinations. For the single modality, text is more effective for depression detection than image and posting time. Besides, image is more important than posting time. The results are reasonable since text can convey more information than image and posting time. Overall, the performance is better when we use more modalities, that is, we can better understand how users feel with more information (Table 7).

5.5.3 Multi-image

On Instagram, each post must have an image and can have up to 10 images. In this subsection, we compare the performance of using only one image and using all the images from each post. Furthermore, when there are multiple images in a post, we compare the results of using minimum, maximum, and average to get the image feature representation of the post. Using average to get the image feature representation of multiple images is better than using a single image since we can better understand how a person feels by considering more information. When there are multiple images, using average to get the image feature representation of each post is better than using maximum and minimum. Using maximum or minimum to get the image feature representation may focus too much or too little on some features so that the results are not as good (Table 8).

Table 9 Comparison of different data collection period

	F1	posts/users
1 month	0.901	6.41
6 months	0.947	26.61
12 months	0.956	42.12

Table 10 Comparison of using single or multiple images

	Precision	Recall	F1
Gui et al.	0.783	0.794	0.788
An et al.	0.842	0.842	0.842
MTAN	**0.885**	**0.931**	**0.908**

5.5.4 Data collection period comparison

For each user, we scraped their posts within one year of the latest posting date. In this subsection, we compare the performance of different data collection periods. As shown in Table 9, the longer the data collection period, the better the performance. When the data collection period is shorter, the collected posts are lesser. This means there may not be enough information for the model to detect the depression of a user.

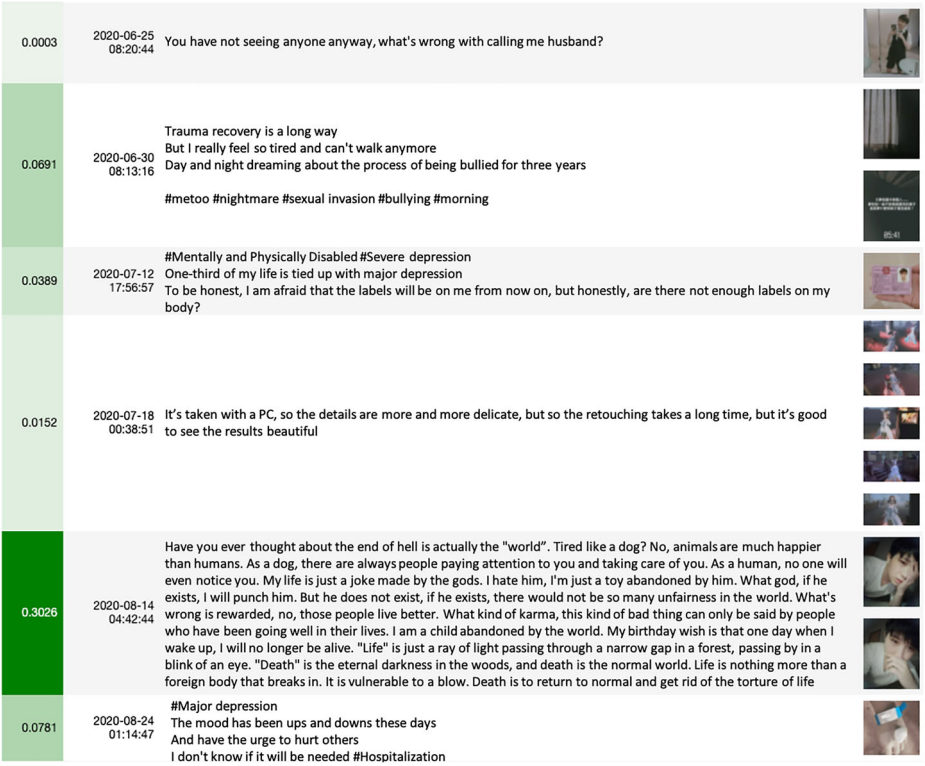

Fig. 8 Posts of the depressive user we found on Instagram. The first column is the attention weight. The second column is the date and time of the post. The third and fourth column is the texts and the images in the post

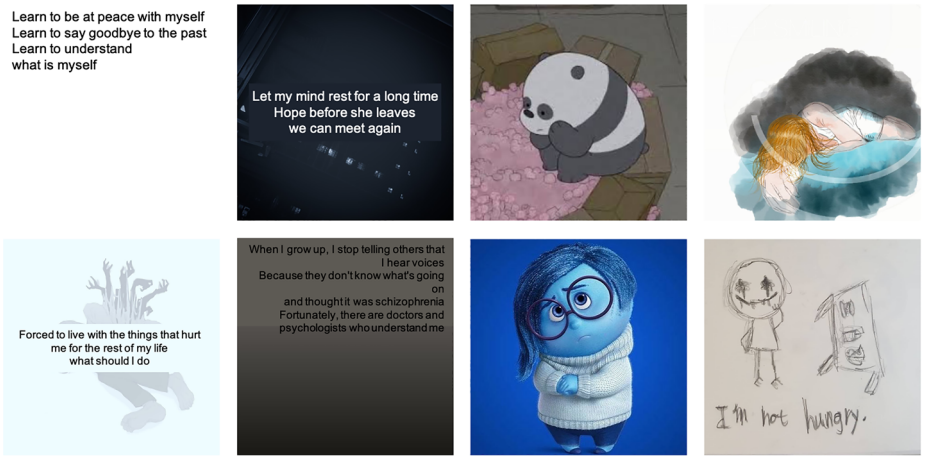

Fig. 9 Examples of images with texts and images with drawings

5.5.5 Performance on twitter dataset

We use the Twitter dataset[3] released by Gui et al. (2019) to test the performance of MTAN. Gui et al. collected the images based on the tweets ids in Shen et al. (2017) dataset to construct a new multimodal dataset. The multimodal dataset has 1,402 depressed users and 1,402 non-depressed users, which has 232,895 and 879,025 posts respectively. Not every tweet has an image, only 22,194 and 64,309 tweets of depressive and non-depressive users have an image. We compared our method with two studies that used this dataset. Gui et al. used reinforcement learning to select the texts and images with depressive moods. The selected texts and images are first averaged respectively and then concatenated to get the predictive result. An et al. (2020) used topic modeling for text and image as two auxiliary tasks to improve the performance of the primary depression detection task. Although Gui et al. managed to select the texts and images with depressive moods, the information of order and time is ignored when averaging the selected features. Instead of selecting or paying more attention to the depressive posts, An et al. considered all the posts in the entire posting history equally, which may be noisy since users have depressive and non-depressive posts. With time-aware mechanism and attention mechanism, our method outperformed Gui et al. and An et al. by 12% and 6.6% in F1-score. The result of depression detection on Instagram is better than the result of depression detection on Twitter. The possible reason is that only 7.8% of the posts on Twitter have images. We can better understand the emotion of the post and the user with the information of images (Table 10).

5.6 Case study

Figure 8 shows an example of a depressive user we found on Instagram. The user has depressive posts and non-depressive posts. Our model can not only detect the user as depressive, but also identify the posts that are depressive. By visualizing the attention weight, we can see that our model gave higher attention weights to the depressive posts.

[3] https://drive.google.com/file/d/11ye00sHFY5re2NOBRKreg-tVbDNrc7Xd/view

5.7 Result analysis

We analyzed the prediction results of the depressed users and found out that the images of the posts with the highest attention weight are mostly grayscale images, images of medicine, and images of blood and wound. ”Pill,” ”pain,” ”die,” ”live,” and ”tired” are the common words that appeared in the highest attention weight posts. We also found out that images with texts and images with drawings are non-obvious image features of the depressed users. Figure 9 shows the examples of images with texts and images with drawings.

6 Conclusion and future work

In this study, we proposed multimodal time-aware attention networks for detecting depression on Instagram. Experiment results demonstrate that incorporating not only textual information but also visual information and posting time is better than using a single modality feature. Furthermore, the performance of depression detection can be improved by incorporating time-aware mechanism and attention mechanism, which outperforms previous work with an F1-score of 95.6%. Visualization of the attention weight from the attention layer shows that our model can effectively select the posts that are important for detecting depression, which provides the interpretability of the model. This indicates the potential of our model to be a reference for psychiatrists to assess the patient; or for users to know more about their mental health condition.

In the process of collecting and analyzing Instagram data, we found that many users tend to post stories instead of posts nowadays. Stories may help to better understand the mental state of the users. The main difference between stories and posts is that stories are available on Instagram only for 24 hours. Therefore, it may take more time to collect stories for depression detection. Furthermore, it is worthwhile to examine different fusion methods to combine multimodal features and other time-aware architectures to see if the model can be further improved.

Acknowledgements This work was partially supported by the Ministry of Science and Technology, ROC (Grant Number: 109-2221-E-468-014-MY3). We thank National Center for High-performance Computing (NCHC) of National Applied Research Laboratories (NARLabs) in Taiwan for providing computational and storage resources.

Data Availability The datasets generated during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of Interests The authors declare that they have no conflict of interest.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., & et al (2016). TensorFlow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX conference on operating systems design and implementation* (pp. 265–283). USENIX Association. https://doi.org/10.5555/3026877.3026899.

American Psychiatric Association (2013). Diagnostic and statistical manual of mental disorders (5th ed). https://doi.org/10.1176/appi.books.9780890425596.

An, M., Wang, J., Li, S., & Zhou, G. (2020). Multimodal topic-enriched auxiliary learning for depression detection. In *Proceedings of the 28th international conference on computational linguistics* (pp. 1078–1089). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.94.

Bahdanau, D., Cho, K. H., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd international conference on learning representations*. arXiv:1409.0473.

Baytas, I. M., Xiao, C., Zhang, X., Wang, F., Jain, A. K., & Zhou, J. (2017). Patient subtyping via time-aware LSTM networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 65–74). https://doi.org/10.1145/3097983.3097997: Association for Computing Machinery.

Beck, A. T., Steer, R. A., & Brown, G.K. (1996). Manual for the Beck Depression Inventory-II. Psychological Corporation.

Chiu, C. Y., Lane, H. Y., Koh, J. L., & Chen, A.L. (2021). *Multimodal depression detection on instagram considering time interval of posts. Journal of intelligent information systems* (Vol. 56, pp. 25–47). Netherlands: Springer. https://doi.org/10.1007/s10844-020-00599-5.

Coppersmith, G., Dredze, M., & Harman, C. (2014). Quantifying mental health signals in Twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality* (pp. 51–60). Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-3207.

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. *Proceedings of the International AAAI Conference on Web and Social Media*, *7*(1), 128–137. https://ojs.aaai.org/index.php/ICWSM/article/view/14432.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Li, F.-F. (2009). Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 248–255). https://doi.org/10.1109/CVPR.2009.5206848.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: human language technologies*, (Vol. 1 pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/n19-1423.

Gui, T., Zhu, L., Zhang, Q., Peng, M., Zhou, X., Ding, K., & Chen, Z. (2019). Cooperative multimodal approach to depression detection in twitter. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 110–117. https://doi.org/10.1609/aaai.v33i01.3301110.

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Huang, Y., Chiang, C.-F., & Chen, A.L. (2019). Predicting Depression Tendency based on Image, Text and Behavior Data from Instagram. In *Proceedings of the 8th international conference on data science technology and applications* (pp. 32–40). https://doi.org/10.5220/0007833600320040.

James, S. L., Abate, D., Abate, K. H., Abay, S. M., Abbafati, C., Abbasi, N., & et al (2018). Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017. *Lancet*, *392*(10159), 1789–1858. https://doi.org/10.1016/S0140-6736(18)32279-7.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of the 3rd international conference on learning representations*. arXiv:1412.6980.

Kroenke, K., Spitzer, R. L., & Williams, J.B. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x.

Mann, P., Paes, A., & Matsushima, E.H. (2020). See and read: Detecting depression symptoms in higher education students using multimodal social media data. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*(1), 440–451. https://ojs.aaai.org/index.php/ICWSM/article/view/7313.

Novak, P. K., Smailović, J., Sluban, B., & Mozetič, I. (2015). Sentiment of emojis. *PLoS ONE*, *10*(12), e0144296. https://doi.org/10.1371/journal.pone.0144296.

Park, M., Cha, C., & Cha, M. (2012). Depressive moods of users portrayed in twitter. In *Proceedings of the 18th ACM international conference on knowledge discovery and data mining* (pp. 1–8).

Radloff, L. S. (1977). The CES-d scale: a Self-Report depression scale for research in the general population. *Applied Psychological Measurement*, *1*(3), 385–401. https://doi.org/10.1177/014662167700100306.

Reece, A. G., & Danforth, C. M. (2017). Instagram photos reveal predictive markers of depression. *EPJ Data Science, 6*(15). https://doi.org/10.1140/epjds/s13688-017-0110-z.

Shen, G., Jia, J., Nie, L., Feng, F., Zhang, C., Hu, T., et al. (2017). Depression detection via harvesting social media: a multimodal dictionary learning solution. In *Proceedings of the twenty-sixth international joint conference on artificial intelligence* (pp. 3838–3844). https://doi.org/10.24963/ijcai.2017/536.

Shen, T., Jia, J., Shen, G., Feng, F., He, X., Luan, H., & et al (2018). Cross-domain depression detection via harvesting social media. In *Proceedings of the twenty-seventh international joint conference on artificial intelligence* (pp. 1611–1617). International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2018/223.

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A.A. (2017). Inception-v4, Inception-ResNet and the impact of residual connections on learning. In *Proceedings of the Thirty-First AAAI conference on artificial intelligence* (pp. 4278–4284). AAAI Press. https://doi.org/10.5555/3298023.3298188.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & et al (2017). Attention is all you need. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 6000–6010). Curran Associates Inc. https://doi.org/10.5555/3295222.3295349.

Wang, P. S., Aguilar-Gaxiola, S., Alonso, J., Angermeyer, M. C., Borges, G., Bromet, E. J., & et al (2007). Use of mental health services for anxiety, mood, and substance disorders in 17 countries in the WHO world mental health surveys. *Lancet*, *370*(9590), 841–850. https://doi.org/10.1016/S0140-6736(07)61414-7.

Wu, M. Y., Shen, C. Y., Wang, E. T., & Chen, A.L. (2020). A deep architecture for depression detection using posting, behavior, and living environment data. *Journal of Intelligent Information Systems*, *54*, 225–244. https://doi.org/10.1007/s10844-018-0533-4.

Yates, A., Cohan, A., & Goharian, N. (2017). Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 2968–2978). Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1322.