Journal of Intelligent Information Systems (2023) 61:497–518
https://doi.org/10.1007/s10844-022-00774-w

The detection of mental health conditions by incorporating external knowledge

Yun Sheng Lin[1] · Liang Kuang Tai[1] · Arbee L.P. Chen[1,2]

Received: 15 September 2022 / Revised: 23 December 2022 / Accepted: 26 December 2022 /
Published online: 18 February 2023

Abstract
Mental health conditions have become a growing problem; it increases the likelihood of premature death for patients, and imposes a high economic burden on the world. However, some studies have shown that if patients are detected and treated early, the social impact and economic costs of mental illness can be reduced. With the popularity of social media, people are sharing their feelings on it, which allows data from social media to be used to study mental health conditions. However, past research had been limited to the optimization of the model or using different types of data available on social media, resulting in models that only rely on data to make decisions. Moreover, people judge things not only by the data collected, but also by background knowledge. Therefore, we considered the diagnostic process of doctors and combined the knowledge of psychological screening tools and diagnostic criteria into the model. In addition, we also tested the effect of combining general knowledge. We retrieve the top m most relevant knowledge segments for each user's post, and then put both into the prediction model. Experimental results show that our method outperforms previous studies, and the F1-score is increased more than 10% in some situations. Moreover, because the knowledge segments are automatically retrieved, our method does not require additional manual labeling, and the knowledge set can be freely adjusted. These show that our method can help detect mental health conditions and can be continuously optimized in practice.

Keywords Mental health condition · Early detection · Social media · Deep learning · Knowledge retrieval

✉ Arbee L.P. Chen
arbee@asia.edu.tw

Yun Sheng Lin
linyunsheng@m108.nthu.edu.tw

Liang Kuang Tai
dlgb429@gmail.com

[1] Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan

[2] Department of Computer Science and Information Engineering, Asia University, Taichung, Taiwan

1 Introduction

According to the World Health Organization (WHO), mental health has become an important indicator of sustainable development (World Health Organization, 2019). Statistics show that people with mental disorders have disproportionately high rates of disability and death. For example, people with schizophrenia and depression are 40% to 60% more likely to die prematurely compared to the general population (World Health Organization, 2021). Through a meta-study of 174 surveys, the number of people suffering from mental illness in the world is estimated at 29.2% (Steel et al., 2014). Worse yet, mental illness is one of the leading causes of disability, driving the global cost of the mental health treatment into trillions of dollars (World Health Organization, 2021; Patel et al., 2018).

Some studies have shown that if mental illness is detected and treated early, the treatment and long-term results will be greatly improved (Bird et al., 2010; Treasure & Russell, 2011). In the long run, being able to early detect mental illness will reduce the impact of mental illness in our society and reduce the economic burden. In WHO's Mental Health Action Plan (World Health Organization, 2021), it suggests to strengthen the research on mental health information systems. We follow this suggestion to focus on early detection of mental illness by analyzing social media data.

Owing to the popularity of the modern Internet, social media is getting closer to people's lives and most people are more and more willing to share their lives through social networks. This allows us to indirectly understand the inner world of people through their posts on social media. With the participation of billions of people, the social media data is large enough to be suitable for deep learning.

Many studies began to collect datasets from social media for analysis or detection of mental health conditions (Coppersmith et al., 2015; MacAvaney et al., 2021; Benton et al., 2017). At the same time, many studies tried to improve the prediction performance by changing the model architecture or considering different types of data (Gui et al., 2019; Shen et al., 2017). Therefore, the performance of the proposed models was often limited by the distribution of the training data, which constrains the generality of the models and also makes the models unable to scale.

Moreover, people judge things not only by the data collected, but also by experiences and background knowledge. This means that in the process of making decisions, people often incorporate relevant knowledge and follow established conventions. Clinically, the doctor or psychotherapist asks the patient questions based on the mental health screening tools to understand their psychological state and symptoms (Butcher et al., 2001; Krug et al., 2008). A final assessment is then made based on the physician's expertise and standard diagnostic criteria (American Psychiatric Association, 2013). Therefore, diagnosing mental illness requires a great deal of knowledge for a precise diagnosis.

To overcome overreliance on data gathered from social media, we refer to human and doctor decision-making processes to incorporate external knowledge into the model. The main idea is to incorporate relevant knowledge from the mental health screening tools and diagnostic criteria into the deep learning model such that a psychological perspective of the model can be provided. As an extension, we also explore the impact of introducing simple common sense into the model. We collected screening tools for the mental health conditions, Wikipedia's mental health-related entries[1,2] and the authoritative book DSM-5 as our

[1] https://en.wikipedia.org/wiki/Template:Mood_disorders
[2] https://en.wikipedia.org/wiki/Template:Mental_disorders

psychological knowledge. The contents from Wiki_dpr[3], which are created by the Hugging Face, are also collected and treated as common sense. Our goal is to study whether external knowledge from psychology or common sense can improve the predictive ability and interpretability of the model. The method of introducing external knowledge has another advantage, that is, the external knowledge can be retrieved automatically and replaced freely. The former means the incorporation of the external knowledge does not need manual annotation, and the relevant knowledge can be found automatically. This will save a lot of labor costs and make our method easier to be widely used. The latter means the contents of the external knowledge can be adjusted freely, which solves the scalability problem for a large-scale model.

Our work includes the following four steps: (1) gather and index the employed external knowledge, (2) incorporate the knowledge into the model to aid in prediction, (3) add an attention layer for determining which posts and knowledge receive more attention, and (4) use a fully connected layer and a sigmoid activation function to predict mental health conditions. Our contributions can be summarized as follows:

- We incorporate relevant external knowledge into the model and the experiment results show that the F1-score of the prediction is increased more than 10% in some situations compared with existing approaches.
- By providing the model with external knowledge from psychology and other fields, humans can better understand what the model has learned, which makes researchers easier to optimize the model.
- This method can be automated, and the external knowledge can be freely adjusted. This minimizes the cost of manual annotation and solves the scalability problem of the model.
- Finally, some statistical guidelines are provided for those who want to adopt our approach to build external knowledge for their model.

The rest of the paper is organized as follows: the related works are reviewed in Section 2, the details of the external knowledge are introduced in Section 3, our approach is described in Section 4, the experiment results are presented in Section 5, and the conclusion is provided in Section 6.

2 Related work

In this section, we first describe the datasets for detecting mental health conditions, and then introduce the past work on the detection of mental health conditions.

2.1 Data collection from social media

With the rise of social media, researchers have begun to use social media data to study mental illness (Park et al., 2012). Moreover, a growing body of research is concentrating on analyzing the copious amounts of text on social media to learn more about mental health conditions (Coppersmith et al., 2015; Birnbaum et al., 2017). However, these studies use manual annotation to label data. Although manual labeling can obtain reliable data, the

[3]https://huggingface.co/datasets/wiki_dpr

number of users from whom the data is collected is limited (Choudhury et al., 2021). Even though crowdsourcing is used to collect and label data, it is still difficult to collect a large amount of user data.

In order to collect larger data, Coppersmith et al. (2014) developed a method to identify self-diagnostic posts on social media by using regular expressions, which was widely used to collect data from the users with mental disorders on social media. Four types of mental health conditions were considered and the Tweets collected were analyzed using corresponding linguistic features and predictive models. Since depressive users tend to express their emotions and even reveal the fact of being diagnosed on social media, labelling users by the method can often achieve a high degree of reliability.

Cohan et al. (2018) extended this approach using Reddit data for a larger number of mental health conditions, and called the dataset Self-reported Mental Health Diagnoses (SMHD). Self-reported diagnoses mean if, for example, "I was diagnosed with depression last year" is found in a post, then this user is considered as a depression patient. The dataset contains data from users with nine different mental health conditions, including depression disorder, attention deficit hyperactivity disorder (ADHD), anxiety disorder, bipolar disorder, post-traumatic stress disorder (PTSD), autism disorder, obsessive-compulsive disorder (OCD), schizophrenia, and eating disorder. Notice that a user may have one or more mental disorders.

For each diagnosed user, nine or more control users were collected according to the following restrictions (Cohan et al., 2018): the number of posts posted by the control user must be between twice and a half of that of the diagnosed user, and the control user must have at least one post on a subreddit where the diagnosed user once posted. It is important to note that these control users cannot have any mental health-related post. Likewise, the diagnosed user is normalized by removing posts containing mental health signals, leaving only general posts in the final dataset, allowing the text analysis to focus on diagnosed user tendencies in general posts. Table 1 shows the statistics of the posts in the two groups of diagnosed users and control users.

Table 1 Average (standard deviation) of the count of posts, tokens for diagnosed and control users in SMHD (Cohan et al., 2018)

	posts		tokens	
condition	per user	total	per post	total
Control	310.0 (157.8)	115,669k	26.2 (48.3)	3,031.6M
Depression	162.2 (84.2)	1272k	45.1 (80.0)	57.4M
ADHD	164.7 (83.6)	872k	46.5 (82.7)	40.5M
Anxiety	159.7 (83.0)	795k	46.4 (83.0)	36.9M
Bipolar	157.6 (82.4)	575k	45.5 (86.5)	26.2M
PTSD	160.7 (84.7)	258k	53.1 (114.0)	13.7M
Autism	168.3 (84.5)	248k	46.5 (82.3)	11.6M
OCD	158.8 (81.4)	203k	46.4 (90.1)	9.4M
Schizophrenia	157.3 (80.5)	123k	49.2 (105.6)	6.1M
Eating	161.4 (81.0)	53k	46.3 (73.7)	2.5M

2.2 Detection of mental health conditions on social media

In order to improve the prediction performance, early studies focused more on the optimization of the model (Jiang et al., 2020; Murarka et al., 2021) or considering different types of data available on social media. Some studies (Choudhury et al., 2021; Coppersmith et al., 2014; Reece & Danforth, 2017) used handcrafted features from different types of data such as number of posts per day, number of faces in a photo, etc. as input for the predictive model. Other studies combined different types of data, such as text and images, to construct a multimodal (Gui et al., 2019; Shen et al., 2017) for the prediction.

However, only relying on social media data to determine whether a user suffers from mental health conditions is less convincing. Past research has focused on adding different types of features, such as the handcrafted features and image features as mentioned above, resulting in the predictive performance of the model limited by the training data. Moreover, because the model parameters cannot be easily changed, the model's generalizability is limited and the model itself unable to scale.

In other fields, researchers solve the above-mentioned problems by incorporating external knowledge into the model. For example, Ghazvininejad et al. (2018) used the memory network to import external knowledge in the conversation generation domain to enable a chatbot to answer questions asked by humans. With the introduction of the external knowledge, the chatbot can answer questions with knowledge not from the training data. For another example, in the question answering domain, Li et al. (2020) used the Word Mover's Distance algorithm to compare the distance between the query and the external knowledge, and put the most matching external knowledge into the model. Li et al. tested the model on the examinations of National Licensed Pharmacist Examination in China, and the results showed that the model can pass the examinations.

Although these methods have shown the effectiveness of incorporating the external knowledge, the method of retrieving relevant knowledge into the model is too straightforward. In recent years, since the excellent feature extraction capabilities of the Pre-trained Language Models (PLMs), the Meta AI team developed a retrieval method based on the high-dimensional feature representation, Dense Passage Retrieval (DPR) (Karpukhin et al., 2020), which greatly increased the accuracy and reliability of the knowledge retrieval. The experiment results show that the retrieval accuracy of DPR not only exceeds the traditionally used BM25 (Robertson & Zaragoza, 2009), but also because it is based on PLMs, its effect can continue to improve with more training. Meta AI team also used DPR to achieve gratifying results in the field of open-domain question answering and generation (Lewis et al., 2020).

Because of DPR's outstanding retrieval ability, we use it to retrieve external knowledge, and then import relevant knowledge and posts into deep learning models for predicting mental health conditions. It is then tested on the SMHD dataset to show the performance of our model.

3 External knowledge

In this section, we first introduce the external knowledge used in our model, and then how we collect this external knowledge, including psychological knowledge and general knowledge.

3.1 Introduction of external knowledge

We consider two types of external knowledge to incorporate into our model. One is the knowledge that a clinical psychologist uses, called *psychological knowledge*, and the other is unconstrained *general knowledge*. By combining these two, we hope to effectively improve the predictive performance of the model and enhance the interpretability of the final results.

3.2 Psychological knowledge

Psychological knowledge is collected from three main sources:

1. Screening tools for the mental health conditions (psychological test questionnaires).
2. Wikipedia's mental health-related entries.
3. DSM-5 (Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (American Psychiatric Association, 2013)).

3.2.1 Screening tools

Screening tools are used to assess an individual's mental health status or to identify signs or symptoms of mental disorders. These tools help clinicians understand the individual's situation to do suitable treatment.

Nine types of screening tools were collected for each mental health condition included in SMHD. Additional screening tools were collected that can be used simultaneously for multiple disorders, such as MMPI (Butcher et al., 2001). In total, there are 10 types of screening tools collected. The collected screening tools are listed in Appendix.

In order to turn the screening tool into model-usable knowledge, it needs to be broken down into *knowledge segments*. We treat each question in the screening tool as a knowledge segment. If the question is too long to fit within the length limit (100 tokens) of a knowledge segment, we divide it into different segments.

For the ten types of screening tools, a total of 50 screening tools with 2556 questions were collected and divided into 2674 knowledge segments as shown in Table 2.

3.2.2 Wikipedia's mental health-related entries

In Wikipedia, related entries are grouped and placed in templates. When collecting Wikipedia data, we found two templates related to mental health conditions, mood disorders, and mental disorders, and scraped all the webpages listed there. These webpages

Table 2 Statistics of the screening tools

Disease	Multiple Condition	ADHD	Anxiety	Autism	Bipolar
# of screening tools	3	5	7	5	4
# of questions	776	236	170	186	144
Disease	Depression	Eating	OCD	PTSD	Schizophrenia
# of screening tools	8	5	5	5	3
# of questions	320	222	193	100	209

contain the description of the diseases, the symptoms, and other information, making the knowledge more comprehensive.

Similar to processing screen tools, Wikipedia data are divided into knowledge segments by sentences to be used by the model. A total of 165 webpages with 22,002 sentences were collected and divided into 22,002 knowledge segments (there is no sentence with more than 100 tokens).

3.2.3 DSM-5

DSM-5 is the Standard diagnostic criteria for mental health conditions. The DSM serves as the principal authority for psychiatric diagnoses in the United States. It contains disease definitions, symptoms, and treatment recommendations, etc. We remove the parts before the preface and after the Appendix, and collect the body part of DSM-5 as psychological knowledge. As above, we use a sentence as a segment and divide the book into many knowledge segments. For long sentences or paragraphs that are difficult to be segmented by automatic tools, we also split them into several segments with 100 tokens as breakpoints.

There are 17329 sentences in total, which are finally split into 17482 knowledge segments. It can be seen that there are not many sentences with more than 100 tokens, and most sentences can completely retain their meanings.

3.3 General knowledge

Wikipedia is a vast online encyclopedia, its content is universal and unrestricted, so we treat it as common sense. The Wikipedia data used was created by the Hugging Face called Wiki_dpr. It covers a wide and large amount of content from Wikipedia.

The text in Wiki_dpr is segmented by 100 tokens, which are split into 21 million segments in total. Wiki_dpr not only contains the text, but also generate the embedding of the knowledge segments by using the same knowledge encoder as ours. The creator establishes a quick search index for these segments, thus users can quickly and accurately find the most relevant knowledge segments.

Since Wiki_dpr contains a wide range of contents, we treat it as common sense to analyze whether the model would perform better if general knowledge was introduced.

4 Mental health conditions detection

In this section, we describe the methods for detecting mental health conditions, including extracting features from the posts and external knowledge segments, finding external knowledge segments related to the posts, and the deep learning model architecture used.

4.1 Feature extraction

To convert text into representative features, a good choice is to represent the text as a high-dimensional vector. In recent years, the rise of Pre-trained Language Models (PLMs) has made extracting features into high-dimensional vectors more universal and effective.

BERT (Devlin et al., 2019) is a transformer-based natural language processing model that is pre-trained on a large corpus and can cover multiple languages at the same time. Because of the self-attention mechanism, the context of the article is taken into account, therefore

the meaning of individual words or the whole article can be accurately expressed. It is one of the most popular pre-trained language models nowadays.

We use BERT as our feature extraction encoder. Both user's posts and knowledge segments are put into respective encoders, and use the output as the final feature representation. In this way, the vector representation of all posts and knowledge segments is obtained, which completes the purpose of the feature extraction.

4.2 Relevant knowledge retrieval

Dense Passage Retrieval (DPR) (Karpukhin et al., 2020) is a tool that has been widely used for knowledge retrieval in recent years. DPR is useful for retrieving relevant content and provides a solution to the scalability problem of large pre-trained models. Its basic concept is to use the excellent feature extraction ability of PLMs to find the correlation between the query and the knowledge such that the knowledge related to the query can be found.

All external knowledge segments are first passed through PLMs to get the feature representation. An index is then built on these knowledge segments such that the relevant knowledge segments can be efficiently retrieved. Next, we pass the query through PLMs to get the feature representation. Finally, the feature representation of all knowledge segments is used to perform Maximum Inner Product Search (MIPS) on the feature representation of the query, and the top m knowledge segments related to the query are obtained.

We treat each post as a query, and let each post find the relevant top m knowledge segments. Then we pass the feature representation of the posts and knowledge segments to the model for the prediction. In the end, the input of the prediction model includes n posts and $m * n$ knowledge segments, where m is a parameter of the number of most relevant knowledge segments.

We train a total of nine binary classification models for the nine corresponding mental health conditions to get nine prediction results. The overall flow is illustrated in Fig. 1.

The details of this method are described as follows.

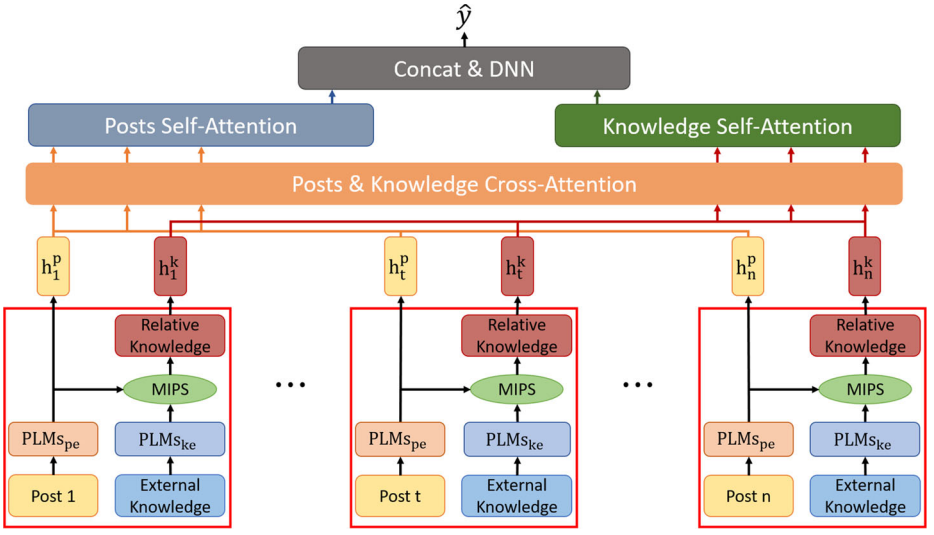

Fig. 1 Architecture for Retrieving External Knowledge and Predicting a Mental Health Condition

Posts $P = \{p_1, p_2, ..., p_t, ..., p_n\}$, where n is the total number of posts and p_t is the t-th post. For each p_t, we use DPR to compute the similarity with all $k_j \in K$, where K represents all knowledge segments:

$$ke(k_j) = PLMs_{ke}(k_j),\ pe(p_t) = PLMs_{pe}(p_t) \tag{1}$$

$$p_\eta(k_j|p_t) \propto exp(ke(k_j)^\intercal pe(p_t)) \tag{2}$$

where $ke(k_j)$ is a feature representation of a knowledge segment produced by the knowledge encoder based on PLMs, and $pe(p_t)$ is a feature representation produced by the post encoder, also based on PLMs. Calculating top-m($p_\eta(\cdot|p_t)$) is a MIPS problem, and we let each p_t find all $p_\eta(k_j|p_t)$), where $k_j \in K$. Then we sort K according to the result to obtain K^t. Lastly, we take top m elements in K^t to get $K^{t,m} = \{k_1^t, k_2^t, ..., k_m^t\}$, which is the top m relevant knowledge segments of p_t. We represent the hidden state of p_t as $h_t^p = pe(p_t)$, and the top-i-th knowledge segment of p_t as $h_t^{k,i} = ke(k_i^t)$.

4.3 Deep learning model architecture

Because different posts and knowledge segments have different contributions to determine mental health conditions, we use an attention mechanism (Vaswani et al., 2017) to let the model pick out posts and knowledge segments that are more important for the prediction.

We use three attention model architectures to test different interactions between posts and knowledge segments:

1. Each post is concatenated with the relevant knowledge segments, then enters the self-attention layer.
2. The posts and knowledge segments enter the self-attention layers separately.
3. The posts and knowledge segments enter the attention layer separately, with cross attention to each other first, followed by the self-attention layer.

The details are as follows.

1. After finding the relevant knowledge segments, let each h_t^p be concatenated with h_t^k to become the t-th hidden state h_t. Put h_1 to h_n into the self-attention mechanism to calculate the attention weight of each hidden state. Then, the output of each hidden state h_t is obtained by a linear combination of the weights. The output goes through a global average pooling layer to obtain the user representation υ.

$$h_t = h_t^p \oplus h_t^k,\ h_t^k = \{h_t^{k,1}, h_t^{k,2}, ..., h_t^{k,m}\} \tag{3}$$

$$\upsilon = pooling(self\_attention(h_1, h_2, ..., h_n)) \tag{4}$$

2. The post and the knowledge segments are then passed to the attention mechanism separately to calculate the attention weights between the posts and between the knowledge segments. After the same linear combination of the weights is done, a global average pooling layer is performed. Then we concatenate the weighted average hidden states of posts and knowledge segments to obtain the user representation υ. This method can be shown after removing the cross attention layer in Fig. 1.

$$h^p = (h_1^p, ..., h_n^p),\ h^k = (h_1^{k,1}, ..., h_1^{k,m}, h_2^{k,1}, ..., h_2^{k,m}, h_n^{k,1}, ..., h_n^{k,m}) \tag{5}$$

$$\upsilon = pooling(self\_attention_p(h^p)) \oplus pooling(self\_attention_k(h^k)) \tag{6}$$

3. The cross attention on the sequences of posts and knowledge segments allows the posts and knowledge segments to interact with each other. Then as in the second method, the

posts and knowledge segments are passed to the attention mechanism separately, followed by the global average pooling layer. Then we concatenate the weighted average hidden states of posts and knowledge segments to get the user representation υ. The complete process is shown in Fig. 1.

$$h_c^p = cross\_attention_{k \to p}(h^k, h^p), h_c^t = cross\_attention_{p \to k}(h^p, h^k) \quad (7)$$

$$\upsilon = pooling(self\_attention_p(h_c^p)) \oplus pooling(self\_attention_k(h_c^k)) \quad (8)$$

After getting the user representation, let it go through the fully connected layer and a sigmoid activation function to get the predicted $\hat{y}$, where $\hat{y}$ is the predictive value of binary classification for each mental health condition:

$$\hat{y} = sigmoid(W\upsilon + b) \quad (9)$$

5 Experiments

In this section, we first give dataset statistics. Then we present the setup of our experiment and analyze the results of the experiment. Next we provide some statistical guidelines of knowledge sources. Finally, we explain the effectiveness of our model with cases.

5.1 Dataset statistics

We use SMHD to validate our method, which has been presented in Section 2.1. It has been divided into training, validation and test sets in equal proportions (Cohan et al., 2018). The statistics are shown in Table 3.

5.2 Experiment setup

Our models are trained with batch sizes of 32 and 8. The optimizer for training the models is Adam with an initial learning rate of 10^{-4}. The loss function used by gradient descent is binary cross entropy. We use Tensorflow2 to implement the models and train the models on NVIDIA Tesla V100.

In terms of feature extraction, we use BERT as our PLMs, and freeze the training parameters of BERT without fine-tuning. The main reason is that the feature extraction ability of the pre-trained BERT is good enough to test the effectiveness of our method. For retrieving relevant knowledge segments, we use a pre-trained encoder trained by Hugging Face as PLMs for our knowledge encoder.[4]

Cohan et al. (2018) selected more than nine control users for each diagnosed user, and mixed all control users and all diagnosed users. In order to fairly compare the experiment results, Sekulic and Strube (2019) set the ratio of the diagnosed users to the control users as 1 : 9, and implemented the benchmarks in Cohan et al. (2018) for a comparison. We follow this setup in our experiments. It is important to note that our way of adjusting the user ratio is different from Sekulic and Strube (2019), where the ratio of the diagnosed users to control users is adjusted through multiple random sampling. We did it by first making predictions for all users and then adjusting the values of the false positive (FP) and true negative (TN). Because the sum of the values of TP and FN is equal to the number of diagnosed users, and

[4] https://huggingface.co/docs/transformers/model_doc/rag

Table 3 Train, Validation, Test Split

	Control	Depress.	ADHD	Anxiety	Bipolar	PTSD	Autism	OCD	Schizo.	Eating
Train	92725	2662	1768	1711	1216	528	479	409	238	104
Val	92421	2574	1747	1593	1182	516	480	477	278	115
Test	94415	2611	1779	1675	1247	558	517	390	267	112

the sum of the values of FP and TN is equal to the number of control users, we calculated the variable x based on the ratio of the diagnosed to control users for each mental health condition to fit the following equation:

$$(TP + FN) : (\frac{FP}{x} + \frac{TN}{x}) = 1 : 9 \tag{10}$$

We then used the values of FP/x and TN/x to calculate the F1 score. Because we make predictions for all users, the results are not biased. In contrast, random sampling has to be done a sufficient number of times to remove bias. Our experiment results are therefore more statistically significant and more representative of the true predictive power of the model.

Furthermore, we only make predictions on 160 posts per user due to the recommendation from Sekulic and Strube (2019). If the user has more than 160 posts, the most recent 160 posts are selected; if it is less than 160 posts, a zero vector is filled.

From Table 3, we see that the number of control users is much larger than that of diagnosed users. To address the problem of data imbalance, we randomly draw diagnosed users and control users into the batch with the same probability during training. This allows all diagnosed and control users to be fairly used in the training.

5.3 Experiment results

We compare with the results of Sekulic and Strube (2019), whose main method is based on the Hierarchical Attention Network (Yang et al., 2016). The basic concept is to use two layers of GRU-based encoders for feature extraction. The attention operation is performed on the feature representations of the posts to achieve the purpose of obtaining the user representation. This study also re-implemented some classic machine learning architectures based on the benchmark models of Cohan et al. (2018), including Logistic Regression, Linear SVM, and Supervised FastText.

To test the impact of different types of knowledge on the model, we divide the knowledge into two categories for the experiments. One is the psychological knowledge specially collected for this task; the other is the addition of general knowledge to the psychological knowledge, called *total knowledge*. The main purpose of the total knowledge experiment is to see if common sense helps model predictions. For each post, we retrieve the top 1 relevant knowledge segment in the first experiment, so each post has one most related knowledge segment.

The method of incorporating external knowledge segments is detailed in Section 4. We train the first two prediction models with knowledge segments from the two categories, and finally obtain four experiment results, as shown in Table 4. It is seen from the experiment results that psychological knowledge is more effective than total knowledge. Therefore, we further test psychological knowledge using a cross-attention mechanism to test whether the interaction of knowledge segments and posts improves model performance.

Table 4 Prediction results of mental health conditions (1 : 9)

	Depress.	ADHD	Anxiety	Bipolar	PTSD	Autism	OCD	Schizo.	Eating
LR	59.00	51.02	62.34	61.87	69.34	**55.57**	**59.49**	56.31	70.71
SVM	58.64	50.08	61.69	61.30	**69.91**	55.35	58.56	**57.43**	**70.91**
FastText	58.38	48.80	60.17	56.53	61.08	49.52	54.16	46.73	63.73
HAN	**68.28**	**64.27**	**69.24**	**67.42**	68.59	53.09	58.51	53.68	63.94
Only_Post	**68.34**	**64.50**	**69.65**	**68.71**	**73.77**	**62.05**	**67.13**	**65.75**	**72.16**
+TK	68.50	61.89	68.84	67.10	72.61	60.59	62.41	63.68	69.46
+PK	68.86	63.05	69.10	69.59	72.19	61.53	65.54	64.90	71.86
+TK +S	69.86	63.45	70.12	68.91	72.17	60.46	65.62	63.72	69.48
+PK +S	**70.73**	65.03	**71.04**	69.70	74.12	61.77	**67.28**	66.26	**74.91**
+PK +C	69.67	**65.59**	70.91	**70.83**	**75.69**	**62.97**	66.80	**68.17**	72.99

Values in bold indicate the highest value

The F1 score cannot measure extremely imbalanced data. If more data is used in the control group, the precision becomes lower with the same model parameter weights because the number of false positives is more likely to be high. To avoid the above situation, we additionally tested the case of the equal number of diagnosed users and control users as shown in Table 5.

Because the previous baselines do not use PLMs, the feature extraction effect for natural language is poor. In order to show that our experiment results are better not only because of PLMs, but also by including external knowledge , we exclude the external knowledge from our model (called Only_Post) as the baseline to evaluate the effect of including the external knowledge for prediction. In Tables 4 and 5, Only_Post represents only the posts part of our model, PK represents psychological knowledge, and TK represents total knowledge. S represents the model that the hidden states of knowledge segments and posts are passed to the attention layer separately, and C represents the model that the cross-attention layer is used. If there is no S and no C, it means that the hidden states of posts and knowledge segments are concatenated before passing to the attention layer.

Table 5 Prediction results of mental health conditions (1 : 1)

	Depress.	ADHD	Anxiety	Bipolar	PTSD	Autism	OCD	Schizo.	Eating
Only_Post	82.74	78.52	83.82	83.13	84.03	74.70	80.29	78.52	89.07
+TK	81.76	76.46	82.55	84.30	82.36	72.93	77.15	75.62	87.30
+PK	82.79	77.58	83.15	83.92	83.52	73.67	77.92	79.66	90.07
+TK +S	84.05	79.34	82.45	85.36	85.97	74.91	79.31	79.36	83.05
+PK +S	84.48	80.70	85.07	85.65	86.37	79.38	82.36	**82.14**	**90.10**
+PK +C	**86.26**	**82.55**	**86.25**	**86.28**	**87.57**	**81.76**	**83.43**	80.79	88.63

Values in bold indicate the highest value

5.3.1 Analysis of the experiment results

It can be seen from Tables 4 and 5 that with the Only_Post model, it performed better than all the baselines made by the previous work. This also coincides with our conjecture that PLMs have very good feature extraction capabilities. We use this as a baseline to analyze the results of adding external knowledge segments.

It can be found that using psychological knowledge outperforms total knowledge. There are two possible reasons. First, the total knowledge has more than 20 million knowledge segments, which makes it difficult to find the commonalities and differences between various kinds of knowledge segments, and makes the predicting difficult. Second, the total knowledge contains more common sense than the psychological knowledge such that it is easier to find some knowledge segments unrelated to the mental health. This leads to inability to find the relevant knowledge segments, leading to poor results.

In terms of the models, passing the hidden states of knowledge segments and posts to the attention layer separately performs better than concatenating them together. The reason is that the importance of knowledge segments and posts are not consistent. The posts with greater attention weight do not necessarily find the important knowledge segments, and vice versa. Therefore, simply concatenating the hidden states of knowledge segments and posts, and passing to the attention layer reduces the predictive power of the model, because it makes model more difficult to find important contents. On the contrary, passing the hidden states of knowledge segments and posts separately through the attention mechanism results in better performance. Because the model can calculate the hidden states of knowledge segments and posts separately, it becomes feasible to find the hidden states of posts and knowledge segments that are important for model prediction. However, separating the hidden states of knowledge segments and posts make them unable to influence each other. We therefore add a cross-attention mechanism to further improve the prediction effect. Finally, we have achieved good results in most of the disease categories. Among these results, schizophrenia showed the greatest improvement with an increase of more than 10% in F1 score. The incorporation of psychological knowledge and the use of the cross attention layer increases the effect by more than 3%.

5.3.2 Extended experiments: More knowledge segments

This section explores the impact of introducing more external knowledge segments on model predictions. We let each post find the top 1, 3, and 5 related knowledge segments (from Psychological Knowledge) and compare the impact on the model. The results are shown in Tables 6 and 7.

It is found that more related knowledge segments may be helpful for model prediction, but not necessarily. While more imported knowledge segments increases the likelihood of acquiring important and relevant knowledge segments, it also creates more noise and makes

Table 6 Experiments of top 1, 3, and 5 related knowledge segments (1 : 9)

	Depress.	ADHD	Anxiety	Bipolar	PTSD	Autism	OCD	Schizo.	Eating
Top 1	70.73	65.03	71.04	69.70	74.12	61.77	67.28	66.26	**74.91**
Top 3	**70.82**	64.98	**71.95**	**70.62**	**75.50**	**62.87**	**68.87**	**66.55**	72.16
Top 5	70.20	**65.55**	71.20	70.29	75.49	62.37	67.07	66.53	72.38

Values in bold indicate the highest value

Table 7 Experiments of top 1, 3, and 5 related knowledge segments (1 : 1)

	Depress.	ADHD	Anxiety	Bipolar	PTSD	Autism	OCD	Schizo.	Eating
Top 1	84.48	80.70	85.07	85.65	86.37	**79.38**	82.36	**82.14**	**90.10**
Top 3	84.11	80.25	**86.32**	86.09	**87.40**	78.66	**82.50**	77.30	88.32
Top 5	**85.20**	**81.19**	85.29	**86.26**	85.10	78.58	82.35	77.13	85.46

Values in bold indicate the highest value

the model harder to focus on the really important knowledge segments. For some mental health conditions with small number of diagnosed users, the more input of knowledge segments, the harder it is to find useful information, and as a result, the more serious problem of overfitting. For example, in eating disorders, the more knowledge segments is introduced, the worse effect results. Therefore, it is not necessarily helpful to introduce more knowledge segments; it depends on the amount of data and the actual situation of the experiment to make a decision.

5.3.3 Co-occurrence of mental health conditions

In this section, we discuss whether the nine mental health conditions predicted by our binary classification model are able to distinguish the multiple disorders of the user, and analyze the comorbidity.

We use the model with the best predictive performance, i.e. Only_Post + PK + C, and analyze the prediction results of all diagnosed users on the nine disorders to study the comorbidities. We made predictions for the nine disorders for each diagnosed user. We used Exact Match Ratio (EMR), a strict metric where each label in a multi-label needs to be exactly correct, and Hamming Loss (HL), a soft metric used to report the average number of

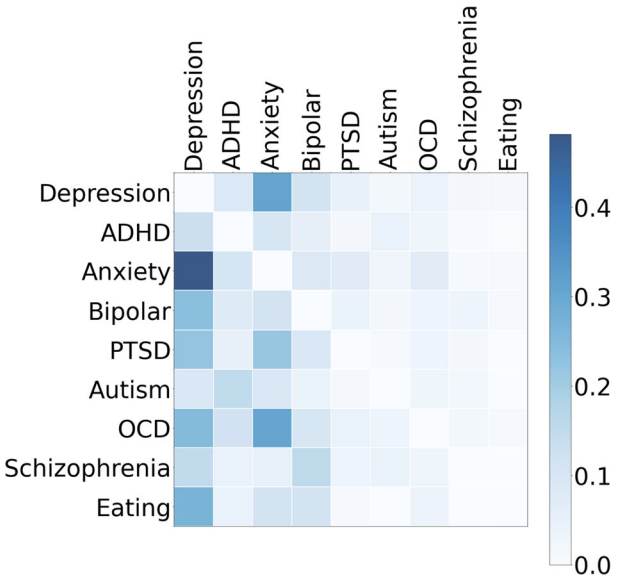

Fig. 2 The relative co-occurrence of the disorder with other disorders in the test dataset

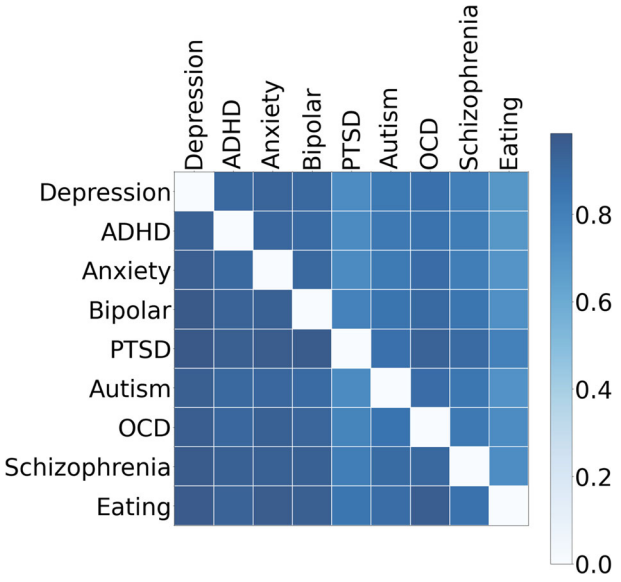

Fig. 3 The relative co-occurrence of the disorder with other disorders in the prediction result

incorrectly predicted class labels. The results show that the EMR is only 1.34%, while the HL result is 61.04%, which shows that our model cannot well distinguish the differences between different disorders.

We compared the comorbidity of the test data, as shown in Fig. 2, with our predicted results, as shown in Fig. 3. It is found from Fig. 3 that as long as the user has a mental health condition, it is almost always predicted to have other mental health conditions. However, the actual situation is not the case. Our model can well detect whether users have a mental health condition, but cannot accurately tell which mental health condition it is.

5.4 Source statistics for psychological knowledge

The psychological knowledge is collected from three sources: Screening tools, Wikipedia (mental health-related entries), and the DSM-5. Table 8 shows a statistical analysis of the knowledge segments retrieved by our method from these three sources.

We search for the top-5 related psychological knowledge segments for all posts and count their sources. It can be seen from Table 8 that for the source distribution of the top 1 5 related knowledge segments, the knowledge segments from Wikipedia (mental health-related entries) are much more often used than the other two. It can be seen from Section 3.1

Table 8 The source distribution of the top 1 5 related knowledge segments

Source	Top 1	Top 2	Top 3	Top 4	Top 5
Screening Tools	8.42%	8.65%	8.76%	8.83%	8.88%
Wikipedia (Mental Health)	**85.07%**	**84.81%**	**84.68%**	**84.60%**	**84.54%**
DSM	6.51%	6.54%	6.56%	6.57%	6.58%

Values in bold indicate the highest value

Table 9 The source distribution of the top ten related knowledge segments that the model pays the most attention to

Conditions \ Sources	Screening tools	Wikipedia (Mental Health)	DSM
Depression	1.99%	**97.93%**	0.09%
ADHD	**49.12%**	40.24%	10.65%
Anxiety	3.18%	**93.03%**	3.80%
Bipolar	29.62%	**65.45%**	4.93%
PTSD	**50.09%**	47.99%	1.92%
Autism	41.69%	**55.65%**	2.66%
OCD	18.14%	**77.76%**	4.10%
Schizophrenia	**52.90%**	39.83%	7.27%
Eating	32.85%	**66.87%**	0.28%

Values in bold indicate the highest value

that DSM-5 occupies the largest proportion of the psychological knowledge, while has the least relevance, and the probability of being retrieved is low. There are two speculative reasons. The first is that the encoder weights of DPR are trained on Wikipedia, which makes the segments from Wikipedia easier to be retrieved. The second is that DSM-5 is a book, compared with the screening tools and Wikipedia (Mental Health), it usually uses abstract or high-level sentences, which leads to a difference from the colloquial posts.

In order to understand which sources of knowledge segments are more important to predict mental health conditions, we analyze the model, Only_Post + PK + S, and make statistics on the sources of the top ten related knowledge segments that the model pays the most attention to. As can be seen from Table 9, DSM-5 is again the least important source compared with screening tools and Wikipedia (Mental Health).

From Tables 8 and 9, it can be concluded that the screening tools and Wikipedia (Mental Health) are very important to the model. If we can increase the amount of data for both, there is a good chance that the model will perform even better.

5.5 Case study

We show in this section why the introduction of external knowledge segments can increase model prediction and improve interpretability. The model used is Only_Post + PK + S.

Because of the user's privacy and data usage agreement, we paraphrased the post. The post attention parts present the three posts with the highest attention weights. The knowledge attention parts present the knowledge segment with the highest attention weight, and the post that retrieved the knowledge segment.

Table 10 shows that the predictions are wrong in the baseline model, but correct after including external knowledge segments. Based on the content of the post, it is difficult to determine that this is a depressed patient. However, from the attentional weight of the external knowledge segments, it is found that patients have a tendency to impulse purchase, making the correct judgment of the model. It is worth noting that some past studies(Mueller et al., 2011; Lejoyeux et al., 1997) have found a fairly high correlation between impulse purchase and depression, which is consistent with the knowledge that the models focus on.

Table 11 is also an example of an incorrect baseline prediction, but it becomes correct after including the external knowledge segments. This example is from an autistic patient. It

Table 10 Example of depression

	Post Attention parts
Top 1 ∼ 3	Posts
Top 1	Description: his experience of successfully **quitting drinking** over 2 years ago. He had a family member who was an alcoholic and was moving in the direction of alcoholism himself. He felt that he was not a typical alcoholic, but agreed that **he needed to address the problem and stopped drinking for a week, and felt great about the results.**
Top 2	Description: **he cut himself**, although not bleeding, but painful. And because there is no bleeding, the healing speed is slow. **He thought it was a small thing, but felt pain.**
Top 3	Description: his anger at seeing some people laugh at the work of beginners who are trying to learn. **He thought it was mean and sad to laugh at someone who was studying.**
Knowledge Attention parts (critical knowledge)	
From which post	Description: a time when **he tossed and slept for a game**, and then **bought a game in a place where he would not normally buy a game**, and **paid double the price**. He felt that the game was **not so exciting** for adults, and **because of the extra cost, he felt more emotional.**
knowledge	Impulse control disorder (wiki): Compulsive shopping or buying is characterized by a frequent irresistible urge to shop even if the purchases are not needed or cannot be afforded.

Values in bold indicate information focused on by the models from the experiments

Table 11 Example of autism

Post Attention parts	
Top 1 ∼ 3	Posts
Top 1	Description: one of his experiences with **sleep paralysis**. He saw a witch do something to him, but he couldn't move. It was a long time ago, but he remembered it well.
Top 2	Description: he needed to be supervised in Empirical Research when he was a child. When he was a teenager, he was looking for a partner who could collide with his thoughts. When he was young, he would not take the initiative to help others.
Top 3	Description: he started to be honest with himself, to do what he wanted, and to stop being influenced by others. In addition to studying, let him have a project that he can control.
Knowledge Attention parts (critical knowledge)	
From which post	Description: **Don't call him unless there is something important.** It should be via email or FB.
knowledge	ABC (screening tools): Resists being touched or held

Values in bold indicate information focused on by the models from the experiments

Table 12 Counter example

posts	knowledge
The sex-files	Premature ejaculation (wiki): These motor neurons are located in the thoracolumbar and lumbosacral spinal cord and are activated in a coordinated manner when sufficient sensory input to reach the ejaculatory threshold has entered the central nervous system
!	Drug overdose (wiki): A drug overdose is the ingestion or application of a drug or other substance in quantities much greater than are recommended.
Description: Share the wonderful experience of **catching Pokémon**. Note at the end of the article that his **capture rate was terrible** today, which made a certain Pokémon run away.	ABC (screening tools): Resists being touched or held
Description: He felt that although **Original Promulgator's mother was a victim of domestic violence**, Original Promulgator's mother was still unforgivable for his domestic violence. Finally console the Original Promulgator.	Phobia (wiki): The hippocampus is a horseshoe-shaped structure that plays an essential part in the brain's limbic

Values in bold indicate information focused on by the models from the experiments

is hard to understand why the model made the prediction with the posts. By guessing from the most important knowledge segments, we see that users do not like to contact with other people, which makes one better understand the reason for the prediction.

Table 12 lists some counter-examples that illustrate the inadequacies of current methods. Since our method retrieves knowledge segments for each post, some unimportant posts find noisy knowledge segments during training and testing. This problem of miscited knowledge segments needs to be overcome when incorporating knowledge segments into the model.

6 Conclusion

In this study, we improved the performance of detecting mental health conditions by incorporating psychological knowledge. The experiment results show that our method outperforms previous work, and the F1-score is increased more than 10% in some situations. By the attentional weight of the knowledge segments, our model can find knowledge segments that are important for predicting mental health conditions and improve interpretability by the content of the knowledge segments. This suggests that our model has the potential to be a reference for psychiatrists to assess patients; or to allow users to learn more about their mental health.

Moreover, DPR is an automatable process, and the external knowledge can be adjusted freely, which make our method more likely to be applied in practice. Through the source statistics for knowledge segments, useful sources can be found to improve the performance of the model. We are working on solving the problem of the miscited knowledge segments

and improving the feature extraction capability of the PLMs and DPR encoders, hoping to achieve an even better performance.

Appendix: List of Screen Tools

1. Multiple Condition:
   (a) MMPI-2 (Minnesota Multiphasic Personality Inventory)
   (b) SCL-90 (Symptom Checklist-90)
   (c) PDQ-4+ (The Personality Diagnostic Questionnaire - Version 4)
2. Depression:
   (a) SDS (Zung Self-Rating Depression Scale)
   (b) MDQ (Mood Disorder Questionnaire)
   (c) HCL-32 (Hypomania Check List)
   (d) HRSD (HAMD, HDRS) (Hamilton Rating Scale for Depression)
   (e) QIDS-SR16 (Quick Inventory of Depressive Symptomatology)
   (f) PHQ-9 (Patient Health Questionnaire-9)
   (g) CES-D (Center for Epidemiological Studies Depression)
   (h) BDI (The Beck Depression Inventory)
3. ADHD:
   (a) ASRS-V1.1 (Adult Self-Report Scale)
   (b) ADHD-RS-IV (ADHD Rating Scale-IV)
   (c) SNAP-IV (Swanson, Nolan, and Pelham-IV Questionnaire)
   (d) SWAN (The SWAN* Rating Scale for ADHD)
   (e) VADTRS (Vanderbilt ADHD Diagnostic Teacher Rating Scale)
4. Anxiety:
   (a) SAS (Zung Self-Rating Anxiety Scale)
   (b) HAM-A (Hamilton Anxiety Scale)
   (c) GAD-7 (General Anxiety Disorder-7)
   (d) SCAS (Spence Children's Anxiety Scale)
   (e) K-GSAD-S (Kutcher Generalized Social Anxiety Scale for Adolescents)
   (f) BAI (Beck Anxiety Inventory)
   (g) Burns Anxiety Inventory
5. Bipolar:
   (a) BRMS (Bech-Rafaelsen Mania Scale)
   (b) BDRS (Bipolar Depression Rating Scale)
   (c) YMRS (Young Mania Rating Scale)
   (d) Goldberg Bipolar Screening Test

6. PTSD:
   (a) PSS-I-5 (PTSD Symptom Scale)
   (b) IES-R (Impact of Event Scale – Revised)
   (c) PCL-5 (PTSD Checklist for DSM-5)
   (d) PSS-SR (PTSD Symptom Scale Self-Report Version)
   (e) SPRINT (Short PTSD Rating Interview)
7. Autism:
   (a) ADI-R (Autism Diagnostic Interview-Revised)
   (b) Q CHAT (Quantitative Checklist for Autism in Toddlers)
   (c) AQ (Autism Spectrum Quotient)
   (d) RBQ-2A (Adult Repetitive Behaviours Questionnaire-2)
   (e) ABC (Autism Behavior Checklist)
8. OCD:
   (a) Y-BOCS (Yale–Brown Obsessive Compulsive Scale)
   (b) BOCS (Brief Obsessive-Compulsive Scale)
   (c) CY-BOCS (Children's Yale-Brown Obsessive Compulsive Scale)
   (d) OCI (Obsessive-Compulsive Inventory)
   (e) FAS (Family Accommodation Scale for OCD)
9. Schizophrenia:
   (a) PANSS (Positive and Negative Syndrome Scale)
   (b) SAPS (Scale for the Assessment of Positive Symptoms)
   (c) SANS (scale for the assessment of negative symptoms)
10. Eating:
   (a) EDDS (Eating Disorder Diagnostic Scale)
   (b) BES (Binge Eating Scale)
   (c) CES (Compulsive Eating Scale)
   (d) SEES (Salzburg Emotional Eating Scale)
   (e) DEAS (Disordered Eating Attitude Scale)

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. Thanks to the Georgetown University Information Retrieval Lab for providing SMHD dataset that allows us to validate our method.

Author Contributions We provide methods for incorporating external knowledge to improve early mental health condition detection model.

Funding Not Applicable

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

Declarations

Ethics approval Not Applicable

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

References

American Psychiatric Association (2013). Diagnostic And Statistical Manual Of Mental Disorders, Fifth Edition. American Psychiatric Association. https://doi.org/10.1176/appi.books.9780890425596.

Birnbaum, M. L., Ernala, S. K., Rizvi, A. F., & et a (2017). A collaborative approach to identifying social media markers of schizophrenia by employing machine learning and clinical appraisals. *Journal of Medical Internet Research*, *19*(8), e289. https://doi.org/10.2196/jmir.7956.

Butcher, J. N., Graham, J. R., Ben-Porath, Y. S., & et a (2001). MMPI-2: Manual for administration, scoring, and interpretation. Revised Edition. University of Minnesota Press.

Benton, A., Mitchell, M., & Hovy, D. (2017). Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th conference of the european chapter of the association for computational linguistics*, (Vol. 1 pp. 152–162).

Bird, V., Premkumar, P., Kendall, T., & et al (2010). Early intervention services, cognitive-behavioural therapy and family intervention in early psychosis: systematic review. *The British Journal of Psychiatry*, *197*(5), 350–356. https://doi.org/10.1192/bjp.bp.109.074526.

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). https://doi.org/10.3115/v1/W14-3207.

Coppersmith, G., Dredze, M., Harman, C., & et al. (2015). From ADHD to SAD: Analyzing the language of mental health on twitter through self-reported diagnoses. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: From Linguistic Signal to Clinical Reality* (pp. 1–10). https://doi.org/10.3115/v1/W15-1201.

Cohan, A., Desmet, B., Yates, A., & et al. (2018). SMHD: A large-scale resource for exploring online language usage for multiple mental health conditions. In *Proceedings of the 27th international conference on computational linguistics* (pp. 1485–1497).

Choudhury, M. D., Gamon, M., Counts, S., & et al (2021). Predicting depression via social media. In *Proceedings of the International AAAI conference on web and social media*, (Vol. 7 pp. 128–137), https://doi.org/10.1609/icwsm.v7i1.14432.

Devlin, J., Chang, M. W., Lee, K., & et a (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 volume 1 (Long and Short Papers)* (pp. 4171–4186). https://doi.org/10.18653/v1/N19-1423.

Ghazvininejad, M., Brockett, C., Chang, M.-W., & et al. (2018). A knowledge-grounded neural conversation model. In *Proceedings of the AAAI conference on artificial intelligence, 32(1)*. https://doi.org/10.1609/aaai.v32i1.11977.

Gui, T., Zhu, L., Zhang, Q., & et al (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.

Jiang, Z. P., Levitan, S. I., Zomick, J., & et al. (2020). Detection of mental health from Reddit via deep contextualized representations. In *Proceedings of the 11th international workshop on health text mining and information analysis* (pp. 147–156). https://doi.org/10.18653/v1/2020.louhi-1.16.

Krug, D. A., Arick, J. R., & Almond, P.J. (2008). ASIEP-3: Autism screening instrument for educational planning - Third Edition.

Karpukhin, V., Oguz, B., Min, S., & et al. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 6769–6781). https://doi.org/10.18653/v1/2020.emnlp-main.550.

Li, D., Hu, B., Chen, Q., & et al. (2020). Towards medical machine reading comprehension with structural knowledge and plain text. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 1427–1438). https://doi.org/10.18653/v1/2020.emnlp-main.111.

Lewis, P., Perez, E., Piktus, A., & et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th international conference on neural information processing systems* (pp. 9459–9474). https://doi.org/10.48550/arXiv.2005.11401.

Lejoyeux, M., Tássain, V., Solomon, J., & et a (1997). Study of compulsive buying in depressed patients. *The Journal of Clinical Psychiatry*, *58*(4), 169–173. https://doi.org/10.4088/jcp.v58n0406.

Mueller, A., Mitchell, J. E., Peterson, L. A., & et al (2011). Depression, materialism, and excessive Internet use in relation to compulsive buying. *Comprehensive Psychiatry*, *52*(4), 420–424. https://doi.org/10.1016/j.comppsych.2010.09.001.

MacAvaney, S., Mittu, A., Coppersmith, G., & et al (2021). Community-Level research on suicidality prediction in a secure environment: Overview of the CLPsych 2021 shared task. In *Proceedings of the*

Seventh Workshop on Computational Linguistics and Clinical Psychology Improving Access(CLPsych) (pp. 70–80). https://doi.org/10.18653/v1/2021.clpsych-1.7.

Murarka, A., Radhakrishnan, B., & Ravichandran, S. (2021). Classification of mental illnesses on social media using RoBERTa. In *Proceedings of the 12th international workshop on health text mining and information analysis* (pp. 59–68).

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 (SIGKDD*, (Vol. 2012 pp. 1–8).

Patel, V., Saxena, S., Lund, C., & et a (2018). The Lancet Commission on global mental health and sustainable development. *The Lancet*, *392*(10157), 1553–1598. https://doi.org/10.1016/S0140-6736(18)31612-X.

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.

Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends®in Information Retrieval*, *3*(4), 333–389. https://doi.org/10.1561/1500000019.

Shen, G., Jia, J., Nie, L., & 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 (IJCAI)* (pp. 3838–3844). https://doi.org/10.24963/ijcai.2017/536.

Steel, Z., Marnane, C., Iranpour, C., & et a (2014). The global prevalence of common mental disorders: a systematic review and meta-analysis 1980-2013. *International Journal of Epidemiology*, *43*(2), 476–493. https://doi.org/10.1093/ije/dyu038.

Sekulic, I., & Strube, M. (2019). Adapting deep learning methods for mental health prediction on social media. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT*, (Vol. 2019 pp. 322–327). https://doi.org/10.18653/v1/D19-5542.

Treasure, J., & Russell, G. (2011). The case for early intervention in anorexia nervosa: theoretical exploration of maintaining factors. *The British Journal of Psychiatry*, *199*(1), 5–7. https://doi.org/10.1192/bjp.bp.110.087585.

Vaswani, A., Shazeer, N., Parmar, N., & et al. (2017). Attention is all you need. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 6000–6010). https://doi.org/10.48550/arXiv.1706.03762.

World Health Organization (2019). The WHO special initiative for mental health (2019-2023): universal health coverage for mental health. https://www.who.int/publications/i/item/special-initiative-for-mental-health-(2019-2023).

World Health Organization (2021). Comprehensive mental health action plan 2013–2030. https://www.who.int/publications/i/item/9789240031029.

Yang, Z., Yang, D., Dyer, C., & et al (2016). Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: human language technologies* (pp. 1480–1489). https://doi.org/10.18653/v1/N16-1174.