Psychiatric Scale Guided Risky Post Screening for Early Detection of Depression

Zhiling Zhang, **Siyuan Chen**, **Mengyue Wu*** and **Kenny Q. Zhu***
Shanghai Jiao Tong University
{blmoistawinde, chensiyuan925, mengyuewu}@sjtu.edu.cn, kzhu@cs.sjtu.edu.cn

Abstract

Depression is a prominent health challenge to the world, and early risk detection (ERD) of depression from online posts can be a promising technique for combating the threat. Early depression detection faces the challenge of efficiently tackling streaming data, balancing the tradeoff between timeliness, accuracy and explainability. To tackle these challenges, we propose a psychiatric scale guided risky post screening method that can capture risky posts related to the dimensions defined in clinical depression scales, and providing interpretable diagnostic basis. A Hierarchical Attentional Network equipped with BERT (HAN-BERT) is proposed to further advance explainable predictions. For ERD, we propose an online algorithm based on an evolving queue of risky posts that can significantly reduce the number of model inferences to boost efficiency. Experiments show that our method outperforms the competitive feature-based and neural models under conventional depression detection settings, and achieves simultaneous improvement in both efficacy and efficiency for ERD.

1 Introduction

Depression has been a major health challenge to the world, with over 280 million people affected, according to WHO[1]. Moreover, the COVID-19 pandemic has further deteriorated the situation. Since people are more willing to express their feelings on online social media during this special period[2], depression detection from online posts can be a promising approach to combat the challenge. The conventional setting of depression detection from online posts is to predict whether a user suffers from depression from the whole posting history [Gui *et al.*, 2019; Zogan *et al.*, 2021a]. However, for social networks which update quickly, another setting, *early risk detection* (ERD) [Losada *et al.*, 2017], may have more potential to detect and offer timely help to risky users. An ERD model should access user post one by one sequentially, dynamically update the estimated risk, and make immediate alert once it is confident enough about its prediction. This setting is less explored due to its unique challenges:

First, the ability of classification on streaming data is a requirement of ERD models. This means that the method is better to be an online, incremental algorithm that can update the prediction every time a user sends a post, rather than an offline batch algorithm that only runs once after a long interval. Since traditional ML models do not come with such ability inherently, a typical solution is to naively process the whole posting history for each update [Trotzek *et al.*, 2018]. This method can hardly be efficient enough in practice. For instance, many systems in an ERD competition, eRisk2019, spent several days for computation [Losada *et al.*, 2019].

Moreover, an ERD model should make tradeoff between its timeliness and accuracy. To make an early prediction, the model usually predicts a depression probability after each update, and makes an alert if the probability exceeds certain threshold, and we can tune the threshold to control the latency of prediction [Trotzek *et al.*, 2018]. Leveraging more posts can certainly facilitate higher accuracy, while it also means that the model makes late predictions, and it fails to make alert before the patient's condition deteriorates. To realize the pareto improvement of both objectives, we should also seek improved model structure. Although large pretrained Language models (LMs) like BERT [Devlin *et al.*, 2018] has achieved great success in many classification tasks, they are seldom applied to ERD, as the long posting history make the memory cost and latency prohibitive.

Due to the sensitiveness of depression detection, model explainability is also a vital property. Without proper explanations, it can be hard for users to trust such novel tools and accept these alerts. Since deep learning models are mostly black-boxes, one cannot ascertain whether their predictions are achieved due to robust features, or some spurious clues [Ribeiro *et al.*, 2020]. Traditional ML models can provide global explanations of the prediction (i.e., feature importance) based on features like word counts. However, it is much more preferable if we can make personalized, symptom-based explanations [Mowery *et al.*, 2017] like psy-

*Corresponding authors are supported by NSFC Grants No.91646205 and No.61901265, SJTU-CMBCC Joint Research Scheme, and Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102).

[1]https://www.who.int/news-room/fact-sheets/detail/depression

[2]https://www.statista.com/statistics/1106498/home-media-consumption-coronavirus-worldwide-by-country/

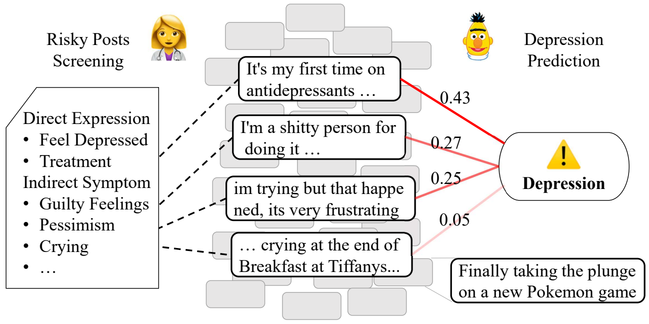

Figure 1: Overview of the system. Depression templates derived from established depression scales are used to screen risky posts, and filter out safer ones (boxes in grey). A hierarchical attentional network further attends to truly important contents (darkness of red lines indicates attention strength) and makes final prediction.

chiatrists to endow higher level of trustworthiness.

Inspired by the psychiatry practice of using clinical scales to screen depression patients, we propose to use depression templates derived from established depression scale [Beck *et al.*, 1996] to screen risky posts. These templates include direct expressions of depressive moods and depression treatments, as well as theory-grounding indirect symptoms like guilty feelings, pessimism and loss of appetite, etc. Only posts highly relevant to these templates will be selected out of the whole posting history, which can greatly reduce the input size, so as to eliminate distractors and improve processing efficiency. A hierarchical network incorporating attention mechanism [Yang *et al.*, 2016] and BERT [Devlin *et al.*, 2018] further aggregates the selected posts of a user, and assigns higher weights to truly important contents for accurate and explainable predictions. The overview of our approach is illustrated in Figure 1. To enable ERD, we also propose an online algorithm based on a risky posts queue evolving with the streaming posts. Experimental results show that the proposed method can achieve SOTA performance in both conventional and ERD settings, and can be even a more efficient ERD solution than simple Logistic Regression models. [3]

Our key contributions are as follows: 1) We propose psychiatry-guided risky post screening to select salient contents for processing, which reduces the input size so as to allow the utilization of large models, and can provide symptom-based interpretations. 2) We leverage hierarchical attentional network with BERT (HAN-BERT) to enhance the model accuracy and explainability. 3) We propose an online algorithm based on an evolving queue of risky posts to tackle ERD, achieving simultaneous improvement in timeliness and accuracy over representative baselines.

2 Methods

For a user U_i with posts $[P_{i,1}, P_{i,2}, ..., P_{i,n}]$ in the activity history, where n is the number of total posts and $P_{i,j}$ is the j-th user-generated post of U_i, the goal of *conventional depression detection* is to predict a binary label $y_i \in \{0, 1\}$ indicating whether the user U_i suffers from depression, given the whole activity history. In contrast, in an *early risk detection* (ERD) setting, the posts come one by one, so that only $[P_{i,1}, P_{i,2}, ..., P_{i,t}]$ is available to the model at the t-th time. The model can make an early prediction of y_i at $t(t \leq n)$ once it is confident enough, such that the prediction can make a good tradeoff between accuracy and earliness (with t as small as possible). Our solutions for both settings are as follows.

2.1 Risky Post Screening

A reddit user typically has hundreds or thousands of posts in the whole activity history. However, since not all user posts are relevant to the detection of depression, retaining all posts may run the risk of introducing distractors that can hinder model performance and efficiency. Therefore, effective post selection strategy can be crucial to the success of ERD.

Intuitively, posts that directly disclose the state of depression or express depression-related symptoms would indicate high depression risks. These intuitions can be reliably captured by psychometrically validated clinical scales. Therefore, we draw inspiration from these scales, and devise **depression templates** to screen risky posts out of the lengthy activity history. Only the posts with highest similarities to these templates will be selected as risky posts for further prediction.

Our depression templates are made up of 2 groups of descriptions. The first group consists of 3 explicit depression-related expressions: "I feel depressed", "I am diagnosed with depression", "I am treating my depression", matching the person's claim of general depressive mood, the diagnosis and the post-diagnosis treatment, respectively. The second group is comprised of descriptions corresponding to dimensions defined in a clinical depression scale. Here we mainly adopt BDI-II, which is one of the most widely used depression measures[4] [Beck *et al.*, 1996]. The scale includes the descriptions of four different intensities for each of the 21 symptoms. For example, "1: I do not feel sad", "2: I feel sad much of the time", "3: I am sad all the time" and "4: I am so sad or unhappy that I can't stand it" for the symptom of sadness. However, we find that current sentence representations have difficulty in capturing such nuanced differences. Therefore, we make slight manual modifications on them to construct a single representative template for each dimension, like "I feel sad" for Sadness. We provide examples in Table 1.

Dimension	Template
Crying	I always cry.
Tiredness	I am too tired to do things.
Self-Dislike	I am disappointed in myself.

Table 1: Example BDI-II dimensions and their corresponding templates.

To measure the similarity between posts and depression templates, we resort to pretrained sentence-transformers

[3]Code, Appendix and the used scale data at: https://github.com/blmoistawinde/scale_early_depress_detect

[4]We also experimented with other scales and combinations of multiple scales, see Appendix for more details.

[Reimers and Gurevych, 2019] to get the sentence representations, and calculate the cosine similarity between each post-template pair. For a post, its most similar template is referred to as its **diagnostic basis**, and their similarity is regarded as the **risk** of the post. The process of **risky post selection** is to select at most K posts with the highest risks out of all posts of a user. Since the procedure consists of only sentence encoding from a pretrained model, and cosine similarity calculations, our method can be more efficient than previous works on post selection that requires costly RL training [Gui *et al.*, 2019]. Moreover, the theoretical underpinning basis for post selection is also more effective than heuristic-based selection such as clustering [Zogan *et al.*, 2021a], as we will validate in the experiments (§3.3).

2.2 Hierarchical Attentional Network

Although depression detection can be formulated as text classification problem, it is different from conventional settings in that the input consists of multiple posts attached with temporal information. We may reformat the input by simply concatenating all posts. However, such representations will lost the time and structural clues at the post level, and also lead to a lengthy sequence. Further, conventional text classification models are lacking in explainability.

To leverage the posting list structure as well as providing post-level explainability, we adopt the framework of Hierarchical Attentional Network (HAN) [Yang *et al.*, 2016] in our model design. The HAN consists of a post encoder and a user encoder.

The post encoder takes the words $\{x_1, x_2, ..., x_L\}$ in a single post, and encode them into a post representation p. Thanks to the pre-step of risky post screening, we are able to take advantage of a large post encoder as opposed to shallow CNN or GRU based structures used in previous works [Yates *et al.*, 2017; Zogan *et al.*, 2021b]. Hence we use a pretrained BERT model as the post encoder and the representation of the [CLS] token as the representation for the whole post. Therefore, the post encoder can be represented as:

$$p = BERT_{[CLS]}(x_1, x_2, ..., x_L) \tag{1}$$

Given the representations of the K risky posts $\{p_1, p_2, ..., p_K\}$, the user encoder models the relations between these posts as well as their chronological order to produce updated contextualized representations of each post $\{p'_1, p'_2, ..., p'_K\}$, and further aggregate these embeddings into one user representation u. Here we utilize a transformer structure to model the posts' relations with self-attention and encode the order with positional embeddings. The updated post representations are further passed to an attentional pooling layer, which learns the weight for each post embedding and perform a weighted sum of them accordingly, to get the final user representation. The attention mechanism can distinguish the contributions of each posts with learned weight so that important posts will have a higher influence on the final prediction. After all, the user encoder can be represented as:

$$p'_1, p'_2, ..., p'_K = Transformer(p_1, p_2, ..., p_K) \tag{2}$$

$$\alpha_k = \frac{exp(Wp'_k + b)}{\sum_{k'=1}^{K} exp(Wp'_{k'} + b)} \tag{3}$$

$$u = \sum_{k=1}^{K} \alpha_k p'_k \tag{4}$$

where W and b is learnable weight matrix and bias term of the linear transformation in the attentional pooling layer. Finally, a linear layer on top of the user representation makes the final binary classification of depression. The whole model is trained with the standard binary cross entropy loss.

With the HAN model above, most of the single post will not exceed BERT length limit. The attention weights can also provide explanations for which post is considered vital in the model's decision.

2.3 Evolving Queue for Early Detection

In an ERD scenario, we need to incrementally make predictions each time a user posts, instead of processing the whole posting history once. This brings the computational challenges of frequent *feature updates* and *model inferences*. For example, a traditional feature-based model would have to recalculate the features like LDA topic distribution [Blei *et al.*, 2003] from the whole activity history after each single update, and then make prediction accordingly. Such frequent recalculations can be even intractable for typical DNN solutions. Although post selection strategies can reduce the computational costs at the *model inference* stage to some extent, the selection stage itself can become a performance bottleneck if it is not efficient enough.

Moreover, the frequent updates may also be sensitive to one single depression-like post and easily produce false positive predictions, while the depressed patients tend to suffer from durative symptoms [Kroenke *et al.*, 2001] as opposed to control users who can also be depressive at some moments. Since the ERD setting does not allow modifying the prediction, once the model makes a confident diagnosis (since we may have already taken action to intervene in the dangerous situation), these false positives cannot be corrected later.

To tackle the above challenges, we further propose an online algorithm based on evolving queue of risky posts to adapt the above methods (§2.1, §2.2) for effective and efficient early detection. First, the Risky Post Screening has already provided an efficient basis for the *feature updates* step[5], as discussed above. However, the *model inferences* remain computationally demanding. We observe that we don't have to make a prediction for each post, as some posts are not very helpful or even misleading for the detection of depression, and these posts are exactly what we would filter out with our screening approach as low risk posts. Therefore, we can make prediction updates only if the new post is considered risky enough, so that the number of model inferences can be substantially reduced. The computational costs can be further contained by limiting the number of posts used in inference, so that only the most risky and recent posts will be included.

[5]We may view the selected posts as "features" in our model, since they are the actual inputs for the prediction model.

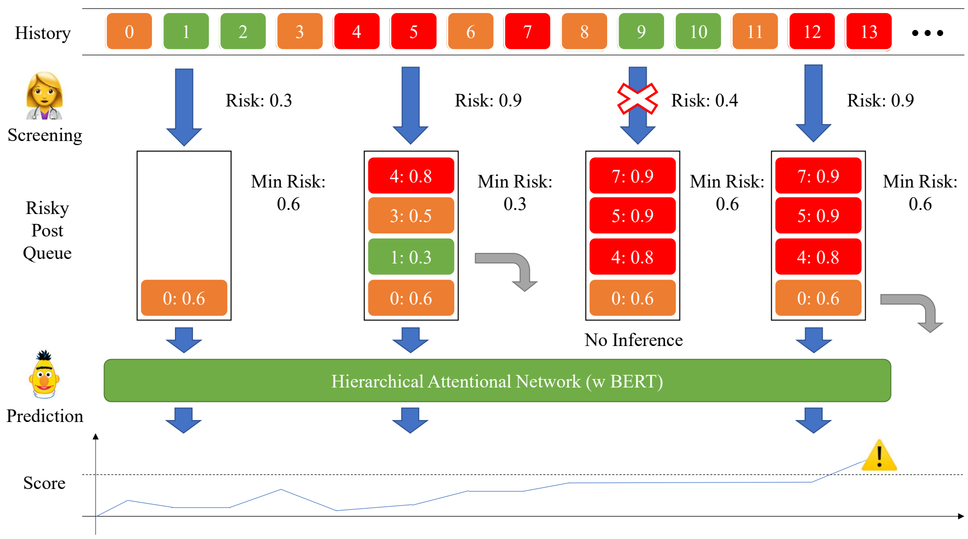

Figure 2: Illustration of the evolving queue for early risk detection.

We implement the above intuition with an evolving queue updated according to post risk. The whole procedure is illustrated in Figure 2. We set the capacity of the queue to K to control the computational costs as well as the number of posts used in training. For each incoming post p, the queue is updated according to the following rules:

1. If the queue is not full, we will add p into the queue no matter how risky it is.
2. If the queue has already been full, we will compare the risk of p with the minimum risk of all posts in the queue. If p is less risky, it will not be included in the queue. Otherwise, the least risky post in the queue will be removed for the insertion of p. The posts in queue will then be sorted in chronological order to align with the model's positional encoding.

The HAN model will make inference only if the queue updates to avoid unnecessary computations. If the model's predicted probability exceeds a predefined threshold, it will report a early alert of depression and stop further calculations.

3 Experiments

In this section, we will first introduce the dataset and the compared methods. We will then further provide our models' results in the conventional depression detection setting and their efficacy and efficiency in ERD settings. We finally illustrate the explainability of the proposed method with examples.

3.1 Dataset

We mainly use the **eRisk2017** dataset [Losada and Crestani, 2016] in our experiment, which is adopted as the benchmark in the ERD task of CLEF 2017 [Losada *et al.*, 2017]. It consists of 137 depressed users and 755 control users and is divided into training/test set with 486/406 users each. The depressed users are identified with patterns like "I was diagnosed with depression", while the control users are those active on depression subreddit but had no depression. The posting year spans from 2007 to 2015. The anchor post for identification is filtered from the dataset. This filtering strategy can prevent the direct information leakage from the self-report, which may prevent the model from learning other indirect depression signals. We also conducted experiments and validated the generalizability of the proposed method on other two datasets with in-domain and cross-domain experiments (see Appendix).

3.2 Competing Methods

We compare our method with several competitive baselines. For traditional machine learning models. **LR** uses TF-IDF features and a logistic regression classifier. **Feature-Rich** utilizes some additional user-based features, including LDA topic distribution [Blei *et al.*, 2003], LIWC features [Pennebaker *et al.*, 2001] and emoticon counts. This is a competitive baseline that has been widely accepted in depression detection works on Facebook, Twitter and Reddit datasets [Eichstaedt *et al.*, 2018; Trotzek *et al.*, 2018; Harrigian *et al.*, 2020].

For neural baselines, we consider both models with large pretrained LM and relatively small models. For small models, we choose the representative **HAN-GRU** model [Zogan *et al.*, 2021b], which adopts a similar HAN structure with GRU as both the user encoder and post encoder. To have a fair competition with large models, it uses the last 1000 posts for classification, which is already a major portion of or the full posting list for many users. For large models, due to their computational cost and length limit, post selection is necessary. Therefore, each method is denoted as a pair of model and selection strategy. For backbone model, we consider the strong pre-trained model **BERT** and the proposed **HAN-BERT**. For post selection strategy, **Heuristic** chooses last posts in user history. **Clus** and **Clus+Abs** are inspired by [Zogan *et al.*, 2021a]. We use sentence-bert [Reimers and Gurevych, 2019] to get post embeddings and run K-means clustering to get the K posts nearest to the cluster center as representative posts

	F1
LR	60.2
Feature-Rich	63
HAN-GRU	61.7
BERT (Clus)	59.6
BERT (Clus+Abs)	52.3
HAN-BERT (Heuristic)	43.2
HAN-BERT (Clus)	62.5
HAN-BERT (Psych)	**70.3**

Table 2: Results on eRisk2017 test set.

(Clus). These posts are further passed to a BART model [Lewis *et al.*, 2020] pretrained on CNN/DM summarization dataset to get an abstractive summary (Clus+Abs). Finally, the proposed screening strategy is denoted as **Psych**.

The basis of BERT and HAN-BERT models are `bert-base-uncased`. The sentence-bert model is `paraphrase-MiniLM-L6-v2`. The number of selected posts is $K = 16$. We train with a batch size of 4, and learning rate of 2e-5. We concatenate the selected posts as input into the BERT baselines. For HAN-BERT models, the user encoder is a 4-layer 8-head transformer encoder. To avoid the influence of randomness, we run each method with 3 different seeds and report the best performance.

3.3 Conventional Setting Results

We first conduct experiments in conventional depression detection setting (Table 2). We can see that BERT (Clus+Abs) performs worse than BERT (Clus), indicating that the abstractive summarization strategy does not necessarily work possibly due to the gap between its pretrained domain (News) and Reddit. HAN-BERT (Clus) outperforms BERT (Clus), showing the effectiveness of the proposed HAN structure. The poor performance of HAN-BERT (Heuristic) highlights the importance of post selection, and none of the traditional post selection methods can outperform the competitive Feature-Rich model with access to all posts. However, with our proposed screening strategy, the HAN-BERT (Psych) model significantly outperforms baselines. HAN-GRU performs worse than HAN-BERT (Psych), suggesting the importance of a strong backbone model.

3.4 Early Detection

We then test model performance in the ERD setting, using the official metrics $ERDE_5$ and $ERDE_{50}$ [Losada *et al.*, 2017]. We also report F1 calculated using the early predictions (report positive if predicted probability over the predefined threshold 0.5, with which all these models are trained). We exclude baselines with unsatisfying performance and preserves LR, Feature-Rich and HAN-GRU. To tackle the item-by-item updates, the baseline models have to recalculate the features and run inference for each new post, while our HAN-BERT with risky post screening can deal this efficiently with the proposed evolving queue algorithm (§2.3). When counting the running time, we accumulate the time costs for all posts no matter if the model makes early predictions to rule out the influence of early false positives. LR and Feature-Rich run on a Linux machine with CPU: E5-2678 (48 cores), while HAN-GRU and HAN-BERT run with a NVIDIA 2080 Ti GPU. Although the comparison is not totally fair, we think it is still a practical setting for real world applications.

Model	$ERDE_5$	$ERDE_{50}$	F1	Time(s)
LR	13.70	8.49	40.5	4710.7
Feature-Rich	12.98	8.39	35.8	7558.4
HAN-GRU	13.30	8.58	40.7	19671.4
HAN-BERT (Psych)	**10.72**	**8.12**	**60.3**	**1330.3**

Table 3: Test results on eRisk2017 in early detection setting. The lower $ERDE_5$ and $ERDE_{50}$, the better model performs early detection.

The results are shown in Table 3. It can be seen that HAN-BERT (Psych) significantly outperforms baselines in ERDE and F1, while also being faster. The reason for the superiority on effectiveness is because baselines produce much more false positives than HAN-BERT due to their sensitivity to single posts. The advantage of efficiency can be mainly attributed to the evolving queue algorithm, which greatly reduced the number of model inference to only 10.41% of all posts. The efficient feature update also helps. Although the sentence encoding must be conducted for all posts, it only costs a small fraction of total time (114.7s out of 1330.3s). The extremely long running time of HAN-GRU further highlights the importance of the proposed algorithm, as we can expect an unaffordable time cost for the even larger BERT-based models without it.

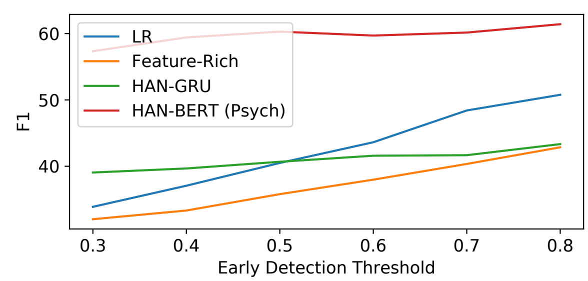

Figure 3: Effect of threshold on early detection F1.

We can adjust the detection latency t by tuning the threshold and balance the tradeoff between precision and recall. Therefore, we hypothesize that model performance can be improved with varied threshold. We tune the threshold from 0.3 to 0.8, and check the changes in their early detection F1. This will run the risk of overfitting on the test set, but allow us to explore the best possible performance. As is shown in Figure 3, the performance of baseline systems can improve by changing the threshold, but still fall behind HAN-BERT (Psych). Moreover, the performance of HAN-BERT (Psych) is not sensitive to threshold, so we may deploy it more comfortably without concerns on threshold tuning.

3.5 Qualitative Example

We provide a concrete example in Table 4 to analyze the behaviors of HAN-BERT (Psych) in detail. From the column of attention weight, we can see that posts with strong depression indicators (e.g. antidepressants, internalizing feelings,

Attention	Post	Diagnostic Basis
0.202	It sucks that the Citalopram didn't work for you but glad to hear your other meds are helping. It's my first time on antidepressants so I didn't know what are their side effects.	**Treatment** Diagnosis Changes in Appetite
0.118	Thanks! :) Sometimes it's really good to actually get the words out of me rather than internalising my feelings.	Concentration Difficulty Loss of Pleasure **Self-Dislike**
0.048	Glad to know :) just glad I'm not working for the next couple of weeks. Feel like I'm on a different planet haha.	**Tiredness** Restlessness Concentration Difficulty
0.021	Some films or TV shows. I remember watching ... The worst part was I'd already been laughed at by my mum for crying at the end of Breakfast at Tiffanys (who leaves a cat out in the rain like that?).	Sadness **Crying** Depressed Mood

Table 4: Example posting list (4 selected out of all 16 posts) of a user with depression with their attention weight in HAN and diagnostic basis according to top 3 cosine similarity (reasonable ones highlighted in bold).

see Row 1, 2) received much higher attention than a uniform baseline (1/16 = 0.0625), while posts with no evident signals of depression or even with a positive emotion (Row 3, 4) received low attention. This suggests the usefulness of the attention weight as an explanation for model prediction. The diagnostic bases decided by the cosine similarity between post and depression templates constitute another type of intuitive explanation. The top 3 diagnostic bases can usually capture the conventional depression behaviors that the post may indicate, which may act as a convincing interpretation in its clinical applications. However, we noticed that sometimes the similarity model may rank an unreasonable aspect high in the list of bases, such as the "Sadness" for the last positive post. We owe such mistakes to the limitation of sentence representation models, such as not sensitive to negation [Ribeiro *et al.*, 2020]. We expect stronger sentence representation models to alleviate the problem.

4 Related Work

Recently, depression detection has received much attention. Studies include predicting depression diagnosis from clinical interviews [Gratch *et al.*, 2014], medical records [Eichstaedt *et al.*, 2018] and self-reported surveys [Guntuku *et al.*, 2019]. Depression detection on social media is especially promising, as proxy diagnostic signals can be relatively easy to get from self-reports or activities in depression communities [Ernala *et al.*, 2019]. Early attempts by Losada and Crestani[2016] used TF-IDF and Logistic Regression on all user posts for depression detection. Later researchers further incorporate new features like LDA, LIWC dictionary and posting patterns [Trotzek *et al.*, 2018]. For deep learning methods, Yates *et al.*[2017] uses hierarchical CNN to process all the posts of a user at the first level and merge the output at the second level for user-level classification. However, most of them directly use all the user's posts without screening out salient posts, which may negatively affect their accuracy and efficiency.

In terms of model interpretability, traditional feature-based methods are partially explainable on the level of global features. For example, Shen *et al.*[2017] found different behaviors for depressed users in posting time, emotion catharsis, self-awareness and life sharing. However, these methods cannot make user-level explanations as personalized diagnostic basis. Detecting depression from its corresponding symptoms can be a promising approach to improve explainability. The pioneering work of Mowery *et al.*[2017] established an annotation scheme for depressive symptoms and an annotated corpus. However, the annotations are difficult so that the amount of data is not sufficient to train a reliable symptom classifier. Our approach also adopts the idea of explaining depression detection from symptoms. But it identifies symptoms implicitly with similarity matching, and thus can alleviate the requirement for large annotated corpus.

In practice, we also want to identify depression risk as early as possible, as is exemplified by the eRisk competitions [Losada *et al.*, 2019]. The majority of proposed methods can only achieve satisfying performance given almost the whole dataset, and few of them are able to make immediate response to each item update. To reduced the number of required posts, Zogan *et al.*[2021a] uses extractive summarization to extract key posts of a user. However, it relies on K-means clustering to get the summaries, so the model cannot run online as well.

5 Conclusions

In this work, we tackle the problem of ERD of depression detection with a novel, psychiatry-guided method of risky post screening and hierarchical attentional network. The accurate selection of risky posts out of the long user history constitutes a solid foundation for prediction as well as enables the usage of large pretrained language model. Furthermore, our framework can work on ERD scenarios with high efficiency, supported by the proposed evolving queue algorithm, which can greatly reduce the required number of model inferences. Utilizing attention mechanism and depression scales provides our method with strong interpretability in the form of attention weights and diagnostic basis, which we hope can facilitate its further application in online detection as a reliable assistant.

Ethical Statement

The datasets used in this work are either publicly available or used under their corresponding data usage agreement. All posts in examples were de-identified and paraphrased for anonymity. We provide further discussion in Appendix.

References

[Beck *et al.*, 1996] Aaron T Beck, Robert A Steer, and Gregory K Brown. *Beck depression inventory (BDI-II)*. Pearson, 1996.

[Blei *et al.*, 2003] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 2003.

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

[Eichstaedt *et al.*, 2018] Johannes C Eichstaedt, Robert J Smith, Raina M Merchant, Lyle H Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A Asch, and H Andrew Schwartz. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 2018.

[Ernala *et al.*, 2019] Sindhu Kiranmai Ernala, Michael L Birnbaum, Kristin A Candan, Asra F Rizvi, William A Sterling, John M Kane, and Munmun De Choudhury. Methodological gaps in predicting mental health states from social media: triangulating diagnostic signals. In *Proceedings of the 2019 chi conference on human factors in computing systems*, 2019.

[Gratch *et al.*, 2014] Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. The distress analysis interview corpus of human and computer interviews. In *Proc. of LREC*, 2014.

[Gui *et al.*, 2019] Tao Gui, Liang Zhu, Qi Zhang, Minlong Peng, Xu Zhou, Keyu Ding, and Zhigang Chen. Cooperative multimodal approach to depression detection in twitter. In *Proc. of AAAI*, 2019.

[Guntuku *et al.*, 2019] Sharath Chandra Guntuku, Daniel Preotiuc-Pietro, Johannes C Eichstaedt, and Lyle H Ungar. What twitter profile and posted images reveal about depression and anxiety. In *Proc. of AAAI*, 2019.

[Harrigian *et al.*, 2020] Keith Harrigian, Carlos Aguirre, and Mark Dredze. Do models of mental health based on social media data generalize? In *Proc. of EMNLP*, 2020.

[Kroenke *et al.*, 2001] Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 2001.

[Lewis *et al.*, 2020] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proc. of ACL*, 2020.

[Losada and Crestani, 2016] David E Losada and Fabio Crestani. A test collection for research on depression and language use. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, 2016.

[Losada *et al.*, 2017] David E Losada, Fabio Crestani, and Javier Parapar. erisk 2017: Clef lab on early risk prediction on the internet: experimental foundations. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, 2017.

[Losada *et al.*, 2019] David E Losada, Fabio Crestani, and Javier Parapar. Overview of erisk 2019 early risk prediction on the internet. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, 2019.

[Mowery *et al.*, 2017] Danielle Mowery, Hilary Smith, Tyler Cheney, Greg Stoddard, Glen Coppersmith, Craig Bryan, and Mike Conway. Understanding depressive symptoms and psychosocial stressors on twitter: a corpus-based study. *Journal of medical Internet research*, 2017.

[Pennebaker *et al.*, 2001] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 2001.

[Reimers and Gurevych, 2019] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proc. of EMNLP*, 2019.

[Ribeiro *et al.*, 2020] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proc. of ACL*, 2020.

[Shen *et al.*, 2017] Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, and Wenwu Zhu. Depression detection via harvesting social media: A multimodal dictionary learning solution. In *Proc. of IJCAI*, 2017.

[Trotzek *et al.*, 2018] Marcel Trotzek, Sven Koitka, and Christoph M Friedrich. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Transactions on Knowledge and Data Engineering*, 2018.

[Yang *et al.*, 2016] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proc. of ACL*, 2016.

[Yates *et al.*, 2017] Andrew Yates, Arman Cohan, and Nazli Goharian. Depression and self-harm risk assessment in online forums. In *Proc. of EMNLP*, 2017.

[Zogan *et al.*, 2021a] Hamad Zogan, Imran Razzak, Shoaib Jameel, and Guandong Xu. Depressionnet: A novel summarization boosted deep framework for depression detection on social media. *arXiv preprint arXiv:2105.10878*, 2021.

[Zogan *et al.*, 2021b] Hamad Zogan, Imran Razzak, Xianzhi Wang, Shoaib Jameel, and Guandong Xu. Explainable depression detection with multi-modalities using a hybrid deep learning model on social media. *arXiv preprint arXiv:2007.02847*, 2021.