Psychiatric Scale Guided Risky Post Screening for Early Depression Detection (Appendix)

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

A Generalizability Tests

Some studies have found that some existing depression detection models do not generalize well in the presence of distribution gaps among online-collected depression datasets [Harrigian *et al.*, 2020]. For instance, within 3 datasets collected from Reddit [Losada and Crestani, 2016; Yates *et al.*, 2017; Wolohan *et al.*, 2018], there are differences in the included subreddits, the time span of the posting histories, and the approaches to annotate the depressed users. As is suggested by [Ernala *et al.*, 2019], current mental health prediction models tend to overfit on the characteristics of a specific dataset instead of learning what they claim to measure (i.e., a robust disease indicator). Therefore, even if there exists similarities for attempts in domain adaptation, the performance of current models still tend to degrade significantly. This highlights the difficulty for existing models to learn generalizable features.

A.1 Experimental Settings

To test the generalizability of baseline models as well as the proposed method, we conduct experiments under conventional depression detection settings on 3 different datasets that have public availability and wide acceptance in previous works [Losada *et al.*, 2017; Trotzek *et al.*, 2018; Harrigian *et al.*, 2020]. They are all collected from Reddit, but have different topic and label distribution and different content filtering strategy that aims to avoid label leakage and simulate user behaviors with different level of self-disclosure.[1] Besides eRisk2017, we introduce the two newly introduced datasets below.

Topic-Restricted consists of 6960 depressed users and 7683 control users, with a 8/1/1 random training/validation/test split [Wolohan *et al.*, 2018]. Since the original dataset is not released with the paper, we follow the implementation of [Harrigian *et al.*, 2020] to crawl the dataset ourselves in the same way. The depressed users are those who started a thread in depression subreddit while the control users are those who started a thread in AskReddit subreddit. The posting year spans from 2007 to 2021. For filtering, all posts in mental health related subreddits are removed. This stricter filtering strategy may further prevent model from overfitting on mental-health related signals, which can not be observed in depressed users who withhold their psychologic status due to the stigma of depression.

RSDD consists of 9210 depressed users and 107274 control users with a training/validation/test split of 39105/39122/39121, after data cleaning [Yates *et al.*, 2017]. The depressed users are also identified with patterns, but further verified by annotators, while 12 control users are matched to a depressed user to minimize their distance of subreddit distribution. The posting year spans from 2006 to 2016. The filtering is the most strict among 3 datasets in that posts either posted in a mental health related subreddit or contain a depression-related term will be removed. This setting forces the model to learn the indirect signals for depression detection, so that they are more likely to detect the depression from patients with no self-report. However, it may also hinder the model's performance on those who would like to share their experience about depression.

For competing methods, we use the baselines described in §3, except HAN-GRU, BERT (Clus+Abs) for efficiency considerations. We additionally compare different variants of the proposed risky post screening strategy, including **Depress** using only 3 direct templates, **BDI-II** using 21 templates derived from BDI-II, and **Full** leveraging a combination of them (i.e. Psych described in §3). On the two new datasets, we find that a tiny version of BERT[2] is enough to achieve competitive results given the larger data size. We select 64 posts, and train with batch size = 32 and learning rate = 2e-4.

A.2 In-domain Results

The in-domain results on the 2 new datasets are shown in Table 2. BERT (Clus) and HAN-BERT (Clus) don't show competitive performance on Topic-Restricted again while requiring expensive clustering stage, so we don't experiment them on the larger RSDD dataset. The proposed risky post screening based methods show strong performance, and are capable of outperforming both the traditional Feature-Rich (all posts) using only 64 posts and a tiny version of BERT. The orders

*Corresponding authors

[1] Although they label "depression" with different methods, they are all valid proxy signals of the same disease ("Depressive Disorder"), covering overlapping but slightly different subsets of patients. Therefore, the robustness across these depression subsets can indicate the general applicability of a method to some extent.

[2] https://huggingface.co/prajjwal1/bert-tiny

Source→Target	T→E	R→E	E→T	R→T	E→R	T→R	Average
LR	82.6	72.3	68.8	67.3	77.8	52.0	70.1
Feature-rich	85.7	75.3	69.6	73.3	77.6	52.7	72.4
HAN-BERT (Depress)	86.6	82.9	75.8	71.1	82.6	**74.8**	79.0
HAN-BERT (BDI-II)	**87.6**	83.8	74.9	**74.6**	80.4	73.5	79.1
HAN-BERT (Full)	87.4	**85.0**	**77.5**	72.4	**84.4**	72.3	**79.8**

Table 1: Cross-domain experimental results (AUC) between eRisk2017(E), Topic-Restricted(T) and RSDD(R).

	Topic-Restricted	RSDD
LR	69.8	52.1
Feature-Rich	72.0	58
BERT (Clus)	56.7	-
HAN-BERT$_{tiny}$ (Heuristic)	68.0	38.2
HAN-BERT$_{tiny}$ (Clus)	71.9	-
HAN-BERT$_{tiny}$ (Depress)	77.1	**65.4**
HAN-BERT$_{tiny}$ (BDI-II)	**78.9**	60.1
HAN-BERT$_{tiny}$ (Full)	77.1	61.1

Table 2: Test F1 on Topic-Restricted and RSDD dataset.

between these screening methods differ across datasets possibly due to their differences in label distribution and filtering strategy, but their performances are overall competitive.

A.3 Cross-domain Results

To test the cross-domain generalizability of different approaches, we train models on a source dataset and directly test the model on another target dataset for all 6 possible combinations of the 3 datasets. For HAN-BERT models, we use the model trained with 16 selected posts, and also test on 16 posts selected with the same templates. In contrast to in-domain experiments using a fixed probability threshold 0.5 to decide the prediction, we don't apply this in cross-domain tests since the label distribution differs greatly between datasets, so that fixed threshold will lead to poor performance. Instead, we use a threshold-free metric, AUC, to measure the performance of each method.

The results are shown in Table 1. Consistent with in-domain results, we find that Feature-Rich outperforms LR in terms of average domain adaptation AUC. Moreover, the performance of all HAN-BERT models are more robust than baselines, which suggests the generalizability of the proposed method.

We then analyze the results in detail. The performance of LR and Feature-rich degrade significantly in the T→R setting. We hypothesize that the reason lies in the different annotation strategy for depressed and control users. Topic-Restricted treats users who started a thread in depression/AskReddit subreddit as depressed/control users and does not control the similarities between the 2 groups, while RSDD specifically selects control users with similar subreddit distribution as depressed users. Therefore, the baselines may have learned spurious clues about the annotation strategy of Topic-Restricted. For example, we checked the coefficients of the LR model, and found that words like "askreddit" and "redditors" are among the most important features for the decision of control users. In contrast, HAN-BERT models still exhibit satisfying performance, which suggests that they can leverage robust depression indicators.

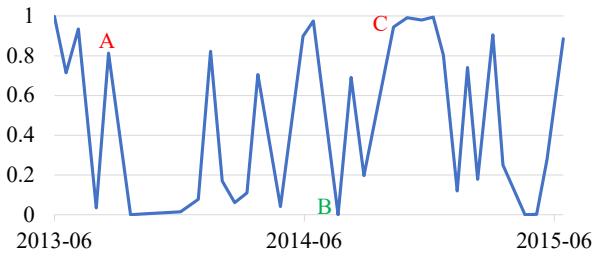
LR and Feature-Rich perform much worse than HAN-BERT models in both direction between RSDD and eRisk2017. These two datasets adopts different filtering strategies, where eRisk2017 preserves most posts while RSDD excludes all posts in mental health related subreddits or containing a depression-related term. The results indicate that HAN-BERT models are less affected by such domain gap.

HAN-BERT models also significantly outperform baselines in the E→T and E→R settings, which shows that they can effectively capture depression signals even with the extremely small eRisk2017 dataset. Comparing all variants of HAN-BERT models, HAN-BERT (Full) shows the best average performance, which indicates the usefulness of combining the theory-guided templates with direct depression indicators in selecting robust features across domains. Therefore, HAN-BERT (Full) can be a preferred choice in real world applications where the target domain distribution is unknown, and we select this variant as the representative of the proposed method.

B Lexical Analysis

As has been shown in many previous works [Shen *et al.*, 2017; Eichstaedt *et al.*, 2018; Wolohan *et al.*, 2018], there are many significant lexical differences between the posts of depressed users and other users, which can be captured by the word frequency of certain categories in LIWC [Pennebaker *et al.*, 2001]. For example, depressed users tend to use more **first person pronouns** (*I*), words expressing **negative emotions** (e.g. *hate, miss, alone*), and words about **health** (e.g. *life, tired, sick*). Such lexical discrepancies do not only exist between the two groups of users, but also within the posts of depressed users themselves. Therefore, we run risky post screening on the posts of depressed users in the eRisk2017 test set, and count the frequency of words in the 3 LIWC categories stated above. We then compare the proportion of these words in selected posts and other posts, and test their differences with two-sided proportion z test. If the selected posts show stronger lexical depression indicators, we can further confirm the helpfulness of risky post screening in capturing reliable depression features.

As is illustrated in Figure 2, there are significant differences between the use of first person words, negative emotions and health-related words in selected (risky) and non-selected posts ($p < 0.001$ for all categories). The difference between risky posts and other posts of depressed users (Negative Emotion: 1.12% vs. 0.74%) can be even bigger than the difference between non-risky posts of depressed users and all



ID	Post
A	When I'm struggling with my shyness and force myself to be confident, I feel a sense of fear. And it f**king hurts if I feel I'm rejected.
B	I have many singers/bands that I'll never tire of.
C	I feel like my depression comes in waves, when I go through good periods and other times I think life is completely pointless.

Figure 1: Predicted depression score by HAN-BERT (Full) along with time for a user in eRisk2017 dataset. We selected 3 time periods on predicted peaks and troughs, and show a representative post in each period.

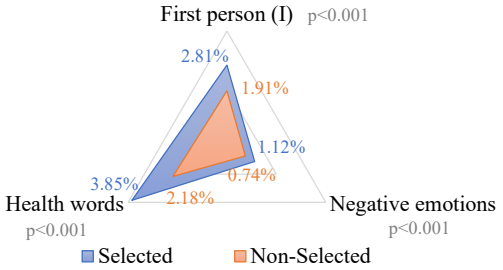


Figure 2: Difference between selected and non-selected posts w.r.t. proportion of words in typical depress related LIWC categories.

posts from non-depressed users (Negative Emotion: 0.74% vs. 0.79%). These findings are similar to those reported in previous literature, which verifies the convergent validity of the proposed method.

C Temporal Analysis

We also demonstrate that our method has the potential to track the fluctuations in depressive mood for depression patients by another example (Figure 1). We produce such curve with the following procedure. First, we group posts according to a 14-day interval. Then we use HAN-BERT (Full) to conduct post screening and depression detection to get a predicted probability for each post group. To produce a smooth curve along time, we design a moving average strategy to derive a more stable depression score from predicted depression probabilities. Suppose the predicted probability of group i is pr_i, and the depression score is s_i. Then we have

$$s_1 = pr_1, s_i = \alpha s_{i-1} + (1-\alpha) pr_{i-1} (i > 1)$$

where $\alpha = max(0, 0.5 * \frac{28 - t_\Delta}{28-1})$, t_Δ is the time interval between the first post of two groups measured in days. Therefore, if two groups are close in time, then the score from the last period will have a higher influence on the next score. Finally, we plot the curve according to the moving-averaged scores.

As we can see in Figure 1. In addition to the accurate detection of depression, the proposed method may also be able to capture the changes in the severity of depression symptoms. The model reported a high risk when the user expressed typical depression symptoms like frustrations (A) or worthlessness (C), and also reported low depression score when the user actually showed interest in things (B), which might indicate a recovery from the symptom of "Loss of Interest". The overall trend is in line with the user's self-report that the depression comes cyclically like waves.

D Hyperparameter Analysis

Here we study the impact of 2 hyperparameters of risky post screening.

Number of Posts We study the effect of post numbers on the Topic-Restricted dataset (Figure 3). It can be seen that all methods can get further improvement given more posts, and scale-based methods can benefit more possibly because more posts can help cover more diverse expressions of depression symptoms, which cannot be fully captured given a small size limit.

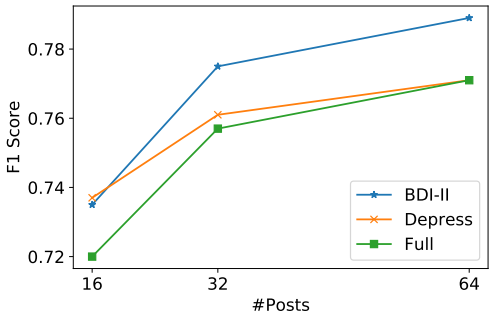


Figure 3: Effect of number of posts on Topic-Restricted.

Different Scales We explore three other commonly-adopted scales besides BDI-II, including HDRS [Hamilton, 1986], CES-D [Radloff, 1977] and PHQ-9 [Kroenke *et al.*, 2001], with similar approaches to rewrite the dimensions into depression templates. We also tried Majority Voting of models paired with different scales. Their performance on eRisk2017 dataset in shown in Table 3. We found that BDI-II is the single best performing scale. The combination of HDRS, BDI-II and PHQ-9 reaches the highest performance for their highly complementary dimensions.

E Depression Templates

Here we provide the templates in detail. We mainly use a combination of 3 direct depression descriptions and the 21 in-

Depression Scale	F1
BDI-II	**70.3**
HDRS	68.0
CES-D	67.9
PHQ-9	67.2
HDRS+BDI-II+PHQ-9	**72.1**
HDRS+BDI-II+CES-D	70.5

Table 3: Test results on eRisk2017 dataset with different depression scales and their combinations.

Dimension	Template
Feeling Depressed	I feel depressed.
Diagnosis	I am diagnosed with depression.
Treatment	I am treating my depression.
Sadness	I feel sad.
Pessimism	I am discouraged about my future.
Past Failure	I always fail.
Loss of Pleasure	I don't get pleasure from things.
Guilty Feelings	I feel quite guilty.
Punishment Feelings	I expected to be punished.
Self-Dislike	I am disappointed in myself.
Self-Criticalness	I always criticize myself for my faults.
Suicidal Thoughts or Wishes	I have thoughts of killing myself.
Crying	I always cry.
Agitation	I am hard to stay still.
Loss of Interest	It's hard to get interested in things.
Indecisiveness	I have trouble making decisions.
Worthlessness	I feel worthless.
Loss of Energy	I don't have energy to do things.
Changes in Sleeping Pattern	I have changes in my sleeping pattern.
Irritability	I am always irritable.
Changes in Appetite	I have changes in my appetite.
Concentration Difficulty	I feel hard to concentrate on things.
Tiredness	I am too tired to do things.
Loss of Interest in Sex	I have lost my interest in sex.

Table 4: The main templates and their corresponding dimensions we used in our experiments, including 3 direct depression descriptions and 21 indirect symptoms derived from BDI-II.

direct symptoms derived from BDI-II (Table 4) [Beck *et al.*, 1996]. As is mentioned in §D, we also experimented other well-known depression scales like HDRS (Table 5) [Hamilton, 1986], CES-D (Table 6) [Radloff, 1977] and PHQ-9 (Table 7) [Kroenke *et al.*, 2001]. The original scales usually contain different descriptions under the same dimension to distinguish different level of intensity or frequency. However, we find that current sentence representations have difficulty in capturing such nuanced differences. We thus condense the descriptions of each dimension into one general template (A few may have more, if there are significant intra-dimension difference).

F Ethical and Broader Impact Statement

This work aims to help people suffering from depression, but have not yet been diagnosed due to the difficulty in receiving clinical help or the stigmatization of the disease. It can be a sensitive topic so it is important to discuss the potential risks and limitations of our work. The proposed method can conduct early depression detection on social media. However, the

I have depressed mood.
I always feel sad.
I feel hopeless.
I feel helpless.
I find myself worthless.
I have feelings of guilty.
I always let people down.
I feel like I should be punished.
I think life is not worth living.
I have thoughts of killing myself.
I tried to suicide.
I have difficulty falling asleep.
I feel restless.
I always wake up during the night.
I have lost my interest in many things.
I decrease time spent in my job.
I find it difficult to concentrate on things.
I can not stay still.
I always worry about small things.
I am irritable.
I feel anxiety.
I have a bad appetite.
I am easy to be tired.
I have less interest in sex.
I suffers from menstrual disturbances.
I worry about my health.
I lose weight dramatically.

Table 5: The templates adapted from the HDRS depression scale.

I am bothered by things that usually don't bother me.
I do not feel like eating.
My appetite is poor.
I feel that I could not shake off the blues even with help -from my family or friends.
I am not just as good as other people.
I have trouble keeping my mind on what I am doing.
I feel depressed.
I feel that everything I did was an effort.
I feel hopeless about the future.
I thought my life had been a failure.
I feel fearful.
My sleep is restless.
I am unhappy.
I talk less than usual.
I feel lonely.
I think people are unfriendly.
It's difficult for me to enjoy life.
I had crying spells.
I feel sad.
I feel that people dislike me.
I could not get 'going'.

Table 6: The templates adapted from the CES-D depression scale.

I have little interest in doing things.
I have little pleasure in doing things.
I always feel down.
I always depressed.
I always hopeless.
I have trouble falling asleep.
I sleep too much.
I feel tired.
I have little energy.
My appetite is poor.
I cannot stop overeating.
I feel bad about myself.
I think myself a failure.
I have let other people down.
I have trouble concentrating on things.
I move much slower than before.
I speak much slower than before.
I have been moving around a lot more than usual.
I think that I would be better off dead.
I have thoughts of hurting myself.

Table 7: The templates adapted from the PHQ-9 depression scale.

performance is far from prefect, so the models' early alerts still require careful examinations from professional practitioners. The proposed method can provide diagnostic bases as explanations, but the diagnostic basis may not precisely matched the actual symptom implied in the post. Therefore, the diagnostic basis should be checked before adoption. Moreover, the datasets are annotated with proxy signals of depression, which may not be representative of the true population of depression patients. In practice, the model should be trained on a more carefully-curated dataset for reliable predictions.

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.

References

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

[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.

[Hamilton, 1986] Max Hamilton. The hamilton rating scale for depression. In *Assessment of depression*. 1986.

[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.

[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.

[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.

[Radloff, 1977] Lenore Sawyer Radloff. The ces-d scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, 1977.

[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.

[Wolohan *et al.*, 2018] JT Wolohan, Misato Hiraga, Atreyee Mukherjee, Zeeshan Ali Sayyed, and Matthew Millard. Detecting linguistic traces of depression in topic-restricted text: Attending to self-stigmatized depression with nlp. In *Proceedings of the First International Workshop on Language Cognition and Computational Models*, 2018.

[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.