Improving the Generalizability of Depression Detection by Leveraging Clinical Questionnaires

Thong Nguyen[1,2], **Andrew Yates**[1,2], **Ayah Zirikly**[3], **Bart Desmet**[4], and **Arman Cohan**[5]

[1]University of Amsterdam, Amsterdam, Netherlands
[2]Max Planck Institute for Informatics, Saarbrücken, Germany
[3]Johns Hopkins University, Baltimore, Maryland
[4]National Institutes of Health, Bethesda, Maryland
[5]Allen Institute for AI, Seattle, WA

Abstract

Automated methods have been widely used to identify and analyze mental health conditions (e.g., depression) from various sources of information, including social media. Yet, deployment of such models in real-world healthcare applications faces challenges including poor out-of-domain generalization and lack of trust in black box models. In this work, we propose approaches for depression detection that are constrained to different degrees by the presence of symptoms described in PHQ9, a questionnaire used by clinicians in the depression screening process. In *dataset-transfer* experiments on three social media datasets, we find that grounding the model in PHQ9's symptoms substantially improves its ability to generalize to out-of-distribution data compared to a standard BERT-based approach. Furthermore, this approach can still perform competitively on in-domain data. These results and our qualitative analyses suggest that grounding model predictions in clinically-relevant symptoms can improve generalizability while producing a model that is easier to inspect.

1 Introduction

Given the significance of mental health as a public health challenge (Brådvik, 2018), much work has investigated approaches for detecting mental health conditions using social media text (Yates et al., 2017; Coppersmith et al., 2018; Shing et al., 2020; Harrigian et al., 2021). Such approaches could be used by at-risk users and their clinicians to monitor behavioral changes (e.g., by monitoring changes in the presence of symptoms related to depression as treatment progresses. These approaches generally rely on datasets consisting of users with self-reported diagnoses (e.g., based on a statement like *"I was just diagnosed with depression"*) for training and evaluation (e.g., Yates et al., 2017; Cohan et al., 2018). Despite promising results on these tasks, related work argues that assessing depression and suicidal behavior is difficult in practical settings and even experienced clinicians frequently struggle to correctly interpret signals (Coppersmith et al., 2018). Furthermore, recent work has found that models trained on particular mental health datasets do not always generalize to others. Harrigian et al. (2020); Ernala et al. (2019) find that systematic, spurious differences between diagnosed and control users can prevent trained models from generalizing to even other, similar social media data. Similarly, outside the mental health domain, recent work reports that neural models often struggle to generalize to data outside their training distribution (Geirhos et al., 2020; D'Amour et al., 2020; Harrigian et al., 2020).

In this work, we explore approaches for constraining the behavior of depression detection methods by the presence of symptoms known to be related to depression, like mood and sleep issues. To do so, we develop nine symptom detection models that correspond to questions present in PHQ9, a screening questionnaire that has been clinically validated and commonly used in practical setting (Kroenke et al., 2001). These questions ask how often the patient has experienced symptoms from nine symptom groups (e.g., how often have you had "*little interest/pleasure in doing things?*").

Grounding depression detection in a trusted diagnostic tool produces several benefits. From the perspective of mental health professionals, the output of such model is inherently more reliable than a black-box model, because classification decisions are based on the presence of specific symptoms in specific posts that can be inspected in order to assess the quality of evidence for a diagnosis. Further, we find this improves the model's ability to generalize, which may be due to limiting its ability to use spurious shortcuts. This strategy is complementary to strategies for reducing temporal and topical artifacts (Harrigian et al., 2020).

Our proposed approach consists of two sim-

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 8446 - 8459

ple yet effective models: a questionnaire model that detects symptoms from PHQ9 and a depression detection model. We instantiate both with a range of methods that are progressively less constrained. At one end of the spectrum, the questionnaire model uses only manually-defined patterns and the depression model makes classification decisions by counting how many times these patterns appear in a user's posts. At the opposite end of the spectrum, there is no explicit questionnaire model and BERT (Devlin et al., 2019a) serves as an unconstrained depression detection model. In between, we relax the questionnaire model by training BERT-based symptom classifiers using the manually-defined patterns, by considering symptom representations rather than counts, and by adding an extra trainable '*other*' symptom.

We find that our constrained models perform competitively compared to a standard unconstrained BERT classifier when trained and evaluated on the same dataset, while additionally providing a model whose behavior can be understood in terms of relevant symptoms in specific posts. However, *dataset-transfer* evaluations demonstrate substantial degradation in BERT's effectiveness. In this setting, our constrained models outperform the unconstrained BERT and show improved generalizability, even across similar datasets.

Our contributions are: *(1)* comprehensive pattern sets for identifying the symptoms in PHQ9 and heuristics for using them to train weakly-supervised symptom classifiers, *(2)* a range of progressively less constrained methods for performing depression detection based on these symptoms, and *(3)* an extensive evaluation of depression detection methods. Our implementation is available online[1].

2 Related Work

Natural language processing methods have been widely used for automatic mental health assessment. To support automated analyses of mental health related language, a variety of datasets have been proposed. Coppersmith et al. (2014) focused on predicting depressed and PTSD users in Twitter, whereas Milne et al. (2016); Shing et al. (2018) and Zirikly et al. (2019) aimed to detect high risk and suicidal users from their ReachOut and Reddit posts, respectively. Yates et al. (RSDD; 2017), Cohan et al. (SMHD; 2018), and Wolohan et al. (2018) investigated identifying depression and other mental health conditions from Reddit. Rich bodies of work in this area focused on studying language use and linguistic styles in depressed users. LIWC (Tausczik and Pennebaker, 2010) has been one of the most popular tools to characterize depression language (Ramirez-Esparza et al., 2008; De Choudhury et al., 2013). Similarly to other NLP domains, the use of contextualized embeddings has improved the performance of classifiers (Jiang et al., 2020; Matero et al., 2019).

Recent work shows that while such NLP models achieve promising results, they have poor generalization to new data platforms and user groups; For example, Harrigian et al. (2020) investigated various factors, including sample size, class imbalance, temporal misalignment (e.g., language dynamic, linguistic norms), deployment latency, and self-disclosure bias that may cause performance degradation when a model is transferred to a new dataset or domain. The issues can occur even when datasets appear to be similar, such as when Reddit-based datasets employ different rules for selecting diagnosed and control users. Another problem is the black-box nature of model predictions which is a major hurdle in deploying AI models in clinical practice (Mullenbach et al., 2018). In this work we aim at reducing this problem by proposing to ground depression assessment in a clinical questionnaire for measuring severity of depression.

Others have considered making predictions more explainable in the mental health domain. Amini and Kosseim (2020) focused on leveraging a user-level attention mechanism for detecting signs of anorexia in social media profiles. Our method differs from theirs in that the explanations are the results of the analysis of the attention weights, while our approaches ground model predictions in a well-established clinical instrument.

Outside of our work, we are aware of two datasets that incorporate questionnaire information such as PHQ9 for identifying depression. The most recent eRisk shared task (Losada et al., 2019) relies on the Beck Depression Inventory (BDI), a 21-item questionnaire that assesses level of depression based on the presence of feelings like sadness, pessimism, etc. Models are built to estimate the user-level BDI score at given time frames. Our approach differs in that we use PHQ9 and evaluate item scores at the post level, which grounds predictions in the presence of clinically-relevant

[1] https://github.com/thongnt99/acl22-depression-phq9

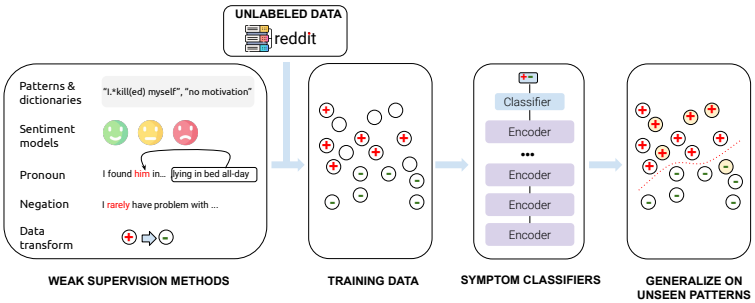

Figure 1: Weakly-supervised questionnaire model

symptoms. In eRisk, a sum of BDI scores is the modeled outcome (corresponding to our baseline pattern-based (threshold) classifier). We use user-level labels for evaluating depression status and evaluate how constraining on PHQ9 symptoms affects the user-level classification performance.

Delahunty et al. (2019) used a deep neural network to predict PHQ4 scores using clinical data that contains patients' PHQ4 scores (Gratch et al., 2014). Our work does not require access to PHQ labeled clinical data, which can be hard to obtain at scale. Furthermore, Delahunty's approach generalizes poorly to social media data. Rinaldi et al. (2020) predict depression based on screening interviews that rely on PHQ9 categories. In their setting, PHQ9 is a channel to retrieve the depressed label, but is not used for explainability. Yadav et al. (2020) propose a multitask learning framework that uses PHQ-9 and figurative language detection as auxiliary tasks. Lee et al. (2021) contemporaneously propose a *micromodel* architecture that they apply to mental health assessment tasks. Our work shares several similarities with this approach, which uses micromodels that are similar to our symptom classifiers (questionnaire models).

3 Methodology

3.1 Pattern-based methods

Our most straightforward and constrained methods are pattern-based classifiers that make classification decisions based on the presence of positive symptom patterns. This method could be decomposed into two components: a questionnaire model and a depression model.

Questionnaire model. The questionnaire model of pattern-based methods is simply a pattern matcher that matches each user post against symptom patterns. It produces a binary pattern matching matrix of size $(num\_post \times 9)$ whose entry at (i, j) is 1 if a match is found between the i^{th} post and any pattern of the j^{th} symptom (question).

Depression model. We implement two variations of the depression model whose input is the pattern matching matrix generated by the previous questionnaire model: a count-based approach and a CNN approach. The count-based approach simply considers whether the number of patterns found in the pattern matrix exceeds a threshold. The CNN approach applies CNN kernels cascaded with a linear layer over the pattern matrix. This approach allows consecutive posts to be weighted differently. In pilot experiments, we also tested a variant that closely mirrors PHQ9 by summing scores over a two-week window; this variant performed worse due to the new temporal requirement that often creates data sparsity within windows.

3.2 Classifier-constrained methods

3.2.1 PHQ9

One drawback of pattern-based classifiers is the inflexibility of pattern matching. The classifier-constrained methods relax the pattern-matching requirement by training a questionnaire model on the weakly-supervised data described in Section 4.2. This results in models that remain grounded in the clinical questionnaire but are capable of generalizing beyond the pattern sets. The PHQ9 architecture is also comprised of a questionnaire model and a depression model.

Questionnaire model. The questionnaire model receives BERT (Devlin et al., 2019b) token embeddings of every post and is trained to predict the answer (positive or negative) for each of the questions in the PHQ9 instrument. This model consists of 9 symptom classifiers, *anhedonia*, *concentration*, *eating*, *fatigue*, *mood*, *psychomotor*, *self-esteem*, *self-harm* and *sleep*, corresponding to the questionnaire's 9 questions. Each symptom classifier is a CNN classifier with a linear layer on top. As illustrated in Figure 1, all symptom clas-

sifiers were separately trained on weakly-labeled data, which we describe in Section 4.2. The questionnaire model's ability to generalize to unseen patterns comes from two sources: BERT embeddings and weakly-labeled symptom data. First, BERT embeddings have been successfully used to transfer knowledge across domains in many NLP applications (Rietzler et al., 2020; Peng et al., 2019; Houlsby et al.). Second, in weakly-labeled data, the background or contextual text around the matched patterns could provide relevant cues, which is a means of generalization. For example, in the text "*now I don't want to do anything. I can't do more than sleep, eat, and watch tv.*", background phrases, such as *"I can't do more..."*, are as useful as the underlined pattern for identifying the symptom *anhedonia*.

Depression model. The depression model predicts whether a user is depressed based on the questionnaire model's output for each post. The questionnaire model's output can be either the final question scores (i.e., symptom scores) or the hidden layers (i.e., symptom vectors) of the 9 sub-models. The former represents each post with a single vector of size 9, which is compact but less informative, while the latter is a larger matrix of size $hidden\_size \times 9$ that preserves more information. Any classification architecture could be used for this depression model. For simplicity, we use a linear classifier on top of features extracted by CNN kernels of various sizes. CNN kernels help summarize symptoms within a sliding window of consecutive posts sorted by timestamp and therefore are a relaxation of the two-week windows considered by the PHQ9 instrument. This relaxation allows more posts to be considered by each CNN kernel, which mitigates the data sparsity problem of the hard two-week window approach.

This depression model is trained using user-level depression labels, and while this model is being trained, the encoder and questionnaire components are frozen. The frozen weights ensure that each questionnaire model does not drift away from its original purpose of detecting symptoms.

3.2.2 PHQ9Plus

PHQ9Plus extends the PHQ9 method by appending an additional symptom (neuron) to the PHQ9 symptoms that form the questionnaire model. This neuron is connected to post embedding and produces a score for every post. Furthermore, we make this additional neuron trainable end-to-end to learn other signals similarly to PHQ9 symptoms. Doing that allows PHQ9Plus to learn from training data other depressive signals in addition to the PHQ9 symptoms. However, in return, it also risks incorporating undesirable shortcuts that harm the model's generalizability.

3.3 Unconstrained method (BERT baseline)

In the previous classifier-constrained methods, depression classifications are constrained by a questionnaire model that determines the presence of symptoms. This is an information bottleneck intended to make the model generalize better. In order to quantify the impact of this bottleneck, we also consider an unconstrained model that replace the questionnaire model in the previous methods by only a BERT encoder. This gives a loose upper bound on the classifier-constrained methods's performance since this approach has access to the raw BERT embeddings and thus can utilize more signals (even spurious ones) than those captured by the questionnaire model.

4 Experiments

We conduct experiments on three datasets; all consist of Reddit social media data but follow different construction methodologies (e.g., identifying depressed users based on a self-report statement vs. based on starting a thread in a support subreddit). In addition to evaluating methods on each dataset, *dataset-transfer* evaluation allows us to evaluate how well methods generalize to similar datasets with different construction methodologies.

4.1 Depression datasets

The three datasets selected for experiments are RSDD (Yates et al., 2017), eRisk2018 (Losada and Crestani, 2016) and TRT (Wolohan et al., 2018). The RSDD (Reddit Self-reported Depression Diagnosis) dataset was constructed from Reddit posts and contains approximately $9,000$ self-reported diagnosed users and $107,000$ matched control users.

eRisk2018 is a smaller dataset of 214 depressed users and $1,493$ control users curated to evaluate the effectiveness of early risk detection on the Internet. Similar to RSDD, the depression group in eRisk2018 was collected based on user self-reports; Here, posts from mental health subreddits were not excluded like in RSDD. Due to the small size of the original training set, which makes the deep learning

approaches unstable, we re-partitioned this dataset to allow more data for training.

Different from RSDD and eRisk2018, TRT (Topic-Restricted-Text) was constructed based on community participation. Specifically, the depressed users were drawn from members of the */r/depression* subreddit, and control users were sampled from the */r/AskReddit* subreddit. Following the construction guideline described in (Wolohan et al., 2018) and discussion with the authors, we re-generated a version of TRT containing $6,805$ depressed users and $57,155$ control users.

On all datasets, we report the F1 score of the positive (i.e., diagnosed) class, and the area under the receiver operating characteristic curve (AUC).

4.2 Questionnaire dataset

The questionnaire model is tasked with classifying if a given post contains a PHQ9 symptom (positive) or not (negative). Given the lack of training data for this task, we collected regular expression patterns and heuristics to construct weakly-supervised training data for each of the symptoms. We describe the process succinctly here and provide additional details in the Appendix. We note that this weakly-supervised data is used only for training.

4.2.1 Positive class

For each question, we prepare a set of positive symptom patterns (e.g., "*can'?t sleep*"). Each pattern set is then matched against a post collection crawled from 127 mental-health subreddits[2]. In addition, we also include posts from the SMHD dataset (Cohan et al., 2018), which excludes posts from mental-health subreddits, to diversify the training data. In the labeling step, we select posts containing symptom patterns as positive examples.

While being fast and transparent, pattern matching may produce many false positives (FPs). We used additional heuristics to remove instances of the four most common types of FPs we observed:

Positive sentiment. Posts containing symptom patterns but conveying a positive/happy sentiment.

Conditional clause. Posts describing a symptom hypothesis rather than an experience.

Third-person pronouns. Posts discussing symptoms of other people (e.g., friends, relatives) rather symptoms the user is experiencing.

Negation. Posts containing symptom patterns with negation words (e.g., "*not*", "*never*") preceding.

[2] https://files.pushshift.io/reddit/

4.2.2 Negative class

Identifying hard negative samples is crucial for the quality of the trained classifiers. We use five heuristics to identify and synthesize negative examples for each symptom:

Keywords. Posts that contain keywords (e.g., "*sleep*") related to positive patterns (e.g., "*can't sleep*") but do not match any positive pattern.

Pronouns. Posts synthesized by replacing first-person pronouns (e.g.,"*I*") from the positive examples with third-person pronouns (e.g., "*She*").

Other symptoms. Posts sampled randomly from positive examples of other symptoms (without matching a pattern for the current symptom).

Negation. Posts synthesized by negating symptom patterns in positive examples using hand-crafted mappings (e.g., "*tired*" to "*never tired*").

Positive sentiment. Posts sampled from neutral or positive classes in the Sentiment140 sentiment analysis corpus (Go et al., 2009).

4.3 Experimental setup

We designed experiments to analyze our two main component: the questionnaire and depression models. The setup for these experiments is summarized in Table 1 and specific hyperparameters are described in the Appendix.

Method	Encoder	Symptom REP(*)	DM(*)
Pattern (threshold)	-	Pattern matrix	-
Pattern (CNN)	-	Pattern matrix	CNN
PHQ9 (scores)	BERT	Scores	CNN
PHQ9 (vectors)	BERT	Vectors	CNN
PHQ9Plus	BERT	Scores + other	CNN
Unconstrained (BERT)	BERT	-	CNN

Table 1: Experimental variations. (*) REP: representation; DM: Depression model

4.4 Depression detection results

The results from prior work and our methods on RSDD, TRT, and eRisk are shown in Table 2.

Depression detection results in *dataset-transfer* evaluation are shown in non-gray blocks in Table 2. Unlike standard within-dataset evaluations, this scenario requires methods trained on one dataset to generalize to other (*highly similar*) datasets. While all datasets consider the same social media platform, their dataset construction methodologies differ, and thus, they are likely to contain different dataset artifacts. Unconstraind methods have the flexibility to learn shortcuts induced by these artifacts, which can lead to poor generalization beyond the training corpus. This effect is observed

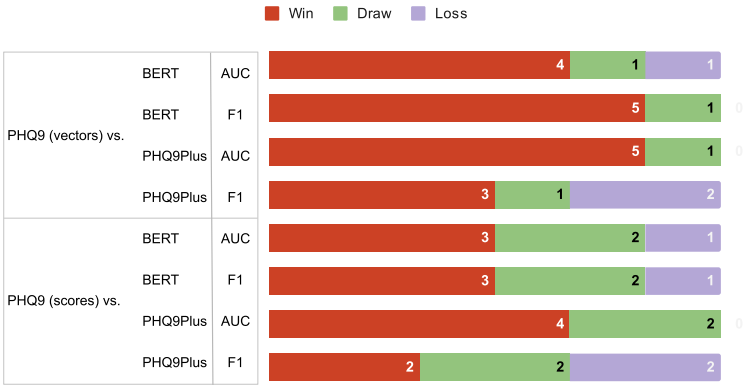

Figure 2: Relative comparison between PHQ9 methods vs. BERT and PHQ9Plus in *dataset-transfer* settings. Win (or Loss): PHQ9 performs significantly better (or worse); Draw: not significant different. (T-test, $\alpha = 0.05$)

Train	Method	Test: eRisk		Test: RSDD		Test: TRT	
		AUC	F1	AUC	F1	AUC	F1
TRT	LIWC+ngram (Wolohan et al., 2018)	-	-	-	-	0.79 ± ∗∗∗	0.73 ± ∗∗∗
	Pattern (threshold)	-	0.38±0.00	-	**0.35±0.00**	-	0.46±0.00
	Pattern (CNN)	0.79±0.01	0.40±0.02	0.71±0.00	0.26±0.01	0.80±0.01	0.51±0.02
	PHQ9 (scores)	0.85±0.01	**0.41±0.01**	**0.78±0.03**	**0.35±0.03**	0.92±0.01	0.64±0.02
	PHQ9 (vectors)	**0.86±0.00**	0.31±0.01	0.73±0.00	0.31±0.00	0.96±0.00	0.77±0.00
	PHQ9Plus	0.80±0.01	0.40±0.07	0.59±0.03	0.21±0.02	0.95±0.00	0.79±0.00
	Unconstrained (BERT)	0.84±0.01	0.15±0.02	0.66 ± 0.03	0.22±0.03	**0.98±0.00**	**0.82±0.00**
RSDD	CNN(400) (Yates et al., 2017)	-	-	-	0.51 ± ∗∗∗	-	-
	Pattern (threshold)	-	0.38±0.00	-	0.35±0.00	-	0.46±0.00
	Pattern (CNN)	0.79±0.01	0.47±0.00	0.74±0.01	0.36±0.02	0.79±0.00	0.39±0.01
	PHQ9 (scores)	0.80±0.01	0.43±0.01	0.85±0.00	0.47±0.01	0.82±0.00	0.46±0.00
	PHQ9 (vectors)	0.81±0.01	0.46±0.01	0.85±0.00	0.49±0.01	**0.86±0.00**	**0.52±0.00**
	PHQ9Plus	0.81±0.03	**0.49±0.00**	**0.86±0.02**	**0.55±0.00**	0.82±0.00	0.49±0.00
	Unconstrained (BERT)	**0.84±0.01**	0.44±0.02	**0.86±0.00**	0.53±0.01	0.82±0.00	0.47±0.00
eRisk	Pattern (threshold)	-	0.40±0.00	-	0.32±0.00	-	0.44±0.00
	Pattern (CNN)	0.80±0.00	0.43±0.01	0.73±0.01	0.31±0.01	0.79±0.00	0.47±0.01
	PHQ9 (scores)	0.87±0.00	0.54±0.02	0.81±0.01	0.38±0.00	**0.90±0.00**	**0.56±0.01**
	PHQ9 (vectors)	0.88±0.00	0.55±0.00	**0.82±0.00**	**0.39±0.01**	0.89±0.00	**0.56±0.04**
	PHQ9Plus	0.94±0.00	**0.73±0.03**	0.79±0.01	0.35±0.01	0.84±0.01	0.54±0.02
	Unconstrained (BERT)	**0.95±0.01**	0.71±0.03	0.81±0.02	0.36±0.02	0.83±0.01	0.50±0.02

Table 2: Depression detection on RSDD, eRisk and TRT datasets (first lines are prior work's results). Highest scores marked in bold. All of our methods and CNN(400) use only a user's first 400 posts, while other baselines use all posts. Summary with statistical test in Figure 2.

Model	Post 1	Post 2
PHQ9 (scores)	Its too late to improve myself [...] I'll graduate soon, but I feel depressed, I'm overweight, and have low confidence and self-esteem [...] now there's nothing left but working for the rest of my life. no friends, no social life, nothing fun - just work. anhedonia mood self-esteem	I'm so tired of living this life [...] I just want it to end. maybe life is just so unfair and there's no explanation for why things are unfair. all the unfair things just get me frustrated [...] im overweight and I never succeed in losing weight, I fuck up every time I try [...] anhedonia self-harm mood fatigue
PHQ9Plus and BERT	I don't like the way my life is going [...] everyday is pretty much entirely spent in my room, except for several hours at the gym [...] That's also why I want people to like me, so that I'd have people to do cool things with and my days would be less lonely and boring.	So what's life like after college? I have to admit that I'm scared as fuck about it. I'm afraid that there will be no time for fun or socializing, and that I'll always have to act all grown up and professional.

Table 3: A depressed user's two most informative posts found by PHQ9 (scores), PHQ9Plus, and Unconstrained (BERT) models. All posts are paraphrased for anonymity.

in the results of the unconstrained model: as summarized in Figure 2, BERT is outperformed by our PHQ9 (scores, vectors) or even pattern-based methods in many dataset-transfer settings. In terms of F1 and AUC, our two PHQ9 variants generalize better than BERT with only 1 "Loss" at most

over 6 *dataset-transfer* settings, and the number of "Win" always dominates. Compared to BERT, our PHQ9 (vectors) obtains 5 "Win", 1 "Draw", and no "Loss" in terms of F1. Regarding AUC, we only observe 1 "Loss" replacing a "Win". The method using PHQ9 scores generalizes slightly worse than the one using vectors but still performs better than the unconstrained models. For example, when trained on TRT and tested on RSDD, our PHQ9 (scores) method improves over BERT by roughly 59% F1 score and 18% AUC. This behavior may reflect the unusual selection of control users in TRT, where control users are sampled from "r/AskReddit". This may introduce shortcuts (e.g., specific topics or styles) that make BERT vulnerable to the change of testing environment. Our classifier-constrained methods with scores and vectors are designed to avoid the spurious shortcuts present in this setting.

For similar reasons, the extra neuron gives the PHQ9Plus model more freedom to learn shortcuts, leading to inferior generalization than PHQ9 (with both scores and vectors). In addition to generalizing better, the methods with symptom scores can be used to identify evidence in the form of specific posts related to the symptoms in PHQ9, which makes them more trustworthy from the perspective of mental health professionals who can examine the posts to verify that symptoms are present.

On standard within-dataset evaluations (in gray cells), when models are trained and tested on the same corpus, we find that F1 and AUC increase as the models become less constrained, with the standalone BERT model and PH9Plus performing the best on all datasets. However, as previously shown, this performance does not transfer to more realistic dataset-transfer settings. The two pattern-based methods perform worse than the best prior method on each dataset, though they are the easiest to interpret due to the PHQ9 symptom scores associated with each post.

When the patterns are used to train a PHQ9 (scores) model, both F1 and AUC increase substantially, with the largest improvement of 0.13 F1 and 0.12 AUC in TRT. Methods using PHQ9 (vectors) perform slightly better than those using scores, but the latter is easier to interpret since each post is associated with a symptom score. Both perform well in comparison with the baselines despite the fact that they are constrained by the PHQ9 symptoms. The add-on neuron contributes significantly to the in-domain effectiveness of PHQ9Plus, which even outperforms BERT in several settings.

Question	**#Pos/Neg**	**F1**	$\kappa_{(*)}$
Anhedonia	75/25	0.54 ± 0.08	0.70
Concentration	72/28	0.83 ± 0.03	0.91
Eating	53/47	0.87 ± 0.01	0.68
Fatigue	46/63	0.62 ± 0.01	0.66
Mood	57/43	0.64 ± 0.02	0.72
Psychomotor	47/53	0.69 ± 0.01	0.80
Self-esteem	52/48	0.77 ± 0.02	0.80
Self-harm	43/57	0.82 ± 0.02	0.80
Sleep	68/32	0.64 ± 0.04	0.68

Table 4: F1 score on manually-labeled samples over five runs and Cohen's kappa (κ) between annotators.

4.5 Symptom detection results

To quantify the performance of our weakly-supervised questionnaire (symptom) models, we additionally prepared a dataset of 900 samples manually labeled by three annotators. The annotation procedures are described in the Appendix B. The results of our symptom classifiers evaluated on the test sets are shown in Table 4. Overall, our symptom classifiers perform well despite being trained on weak labels. The "*concentration*", "*eating*" and "*self-harm*" classifiers show strong performance, while a lower F1 is observed with the "*anhedonia*", "*mood*" and "*fatigue*" classifiers. Interestingly, we find that the F1 scores of symptom classifiers tend to positively correlate with the annotator's agreement (Pearson $\rho > 0.5$). This suggests the low F1 score in some symptom classifiers, such as *"anhedonia"* and *"fatigue"*, might partly be due to the ambiguity of texts. For example, it is challenging to distinguish between an ordinary bad mood versus a depressive mood. Additionally, in our analysis, we find many wrong predictions where posts use symptom-like language in a more specific context, such as "*I completely lost my interest in him*" or "*I can't concentrate on that movie*". These alone might not indicate a symptom, but recurrence of them might be significant.

4.6 Do symptom classifiers generalize?

To examine whether our symptom classifiers can generalize beyond pattern matching, we split each pattern set into two non-overlapping groups (*g1*, *g2*), which split the original dataset into two exclusive subsets. Because pattern distribution are uneven, the resulting subsets are sometimes imbalanced. We then evaluate our symptom classifiers on two settings (i.e., train on g1 & test on g2, and

train on g2 & test on g1). The results shown in Table 5 show that our symptom classifiers still achieve fairly high F1 scores in both settings. Note that, on some symptoms (e.g., concentration, fatigue), given a small coverage of patterns in *g2*, our models could still achieve good performance compared to models trained on the much larger data covered by *g1*. This suggests that the symptom classifiers can generalize beyond the specific patterns they were trained with.

Symptom	*%g1(*)*	*%g2(*)*	Train *g1* Test *g2*	Train *g2* Test *g1*
Anhedonia	0.59	0.41	0.82 ± 0.04	0.79 ± 0.03
Concentraion	0.86	0.14	0.79 ± 0.03	0.72 ± 0.03
Eating	0.53	0.47	0.76 ± 0.02	0.72 ± 0.04
Fatigue	0.96	0.04	0.76 ± 0.03	0.71 ± 0.02
Mood	0.90	0.10	0.72 ± 0.02	0.66 ± 0.04
Psychomotor	0.80	0.20	0.70 ± 0.02	0.66 ± 0.05
Self-esteem	0.95	0.05	0.70 ± 0.02	0.68 ± 0.02
Self-harm	0.60	0.40	0.68 ± 0.02	0.66 ± 0.01
Sleep	0.59	0.41	0.86 ± 0.03	0.72 ± 0.01

Table 5: F1 evaluating symptom classifiers on different pattern sets. (*) The proportion of the original dataset covered by each pattern group.

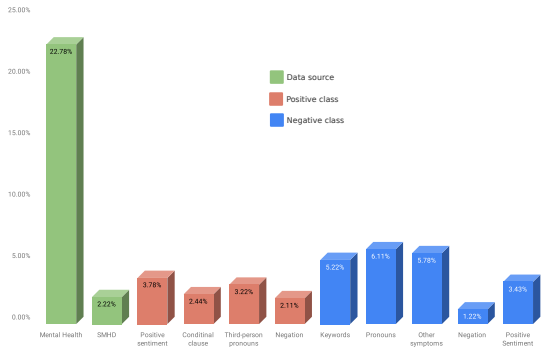

Figure 3: Effect of various factors on symptom detection. Expressed as percentage point difference in F1.

4.7 Effect of labeling methods

In Figure 3, we visualize the effect of various data-construction factors on the performance of symptom classifiers. Regarding the *data source*, data obtained from mental health subreddits has more influence on effectiveness than the more general posts in SMHD. Discarding data from mental health subreddits leads to an average drop of nearly 0.23 F1 score in all symptoms, while the decrease after removing SMHD is 0.02. We attribute the immense contribution of data from mental health subreddits to the fact that mental health is the main topic of discussion in those forums; therefore, pattern matching returns fewer false-positive cases and denser symptoms, resulting in better quality training data.

We further investigate the role of each method to remove FP matches in the *positive class*. For that purpose, we put filtered-out FP examples back into the training data and observe the variation of F1 score on the manually labeled test sets. In general, adding back FP examples filtered by our methods causes a total drop of nearly 0.12 in the averaged F1 score. Among them, instances with positive sentiment cause the highest decrease of roughly 0.04. Posts with third-person pronouns contribute around 0.03, while conditional clause and negation contribute modestly at around 0.02 F1. Similarly, we analyze the effectiveness of methods to weakly annotate the *negative class* by removing each of them from the training data and record the change in F1 score. We find that removing three methods, including keywords, pronouns, and other symptoms, causes a similar drop of roughly 0.06 each. Interestingly, eliminating data with positive sentiment from the *negative class* has a similar effect to adding them to the *positive class*, causing a drop of almost 0.04 F1. The method that changes positive examples to negative examples has the smallest impact on the F1 score (roughly 0.01). Overall, except for the data sources, no single labeling method has a superior impact on the quality of symptom classifiers than other methods.

4.8 Contribution of PHQ9 symptoms

To measure the contribution of a symptom to detecting depression, we remove the corresponding symptom from the model and observe the drop in the F1 score. The results are reported in Table 6. On average, we could see that *"self-harm"*, *"fatigue"*, and *"anhedonia"* are the strongest indicators of depression. Removing them causes a 0.13-0.17 drop in the F1 score. This is in line with the prior finding that suicidal ideation or self-harm is highly correlated with depression (Brådvik, 2018). *"Mood"*, *"psychomotor"*, and *"self-esteem"* contribute moderately to depression detection, with roughly a 0.09 drop in F1 score for each. The remaining three symptoms, including concentration, eating, and sleep, play a less important role in detecting depression, with each contributing around 0.05 to the F1 score.

4.9 Comparison with Few-shot Learner

Recent work has demonstrated that GPT-3 is a strong few-shot learner (Brown et al., 2020). Herein, we are interested in how well our classifier-constrained methods compare to the GPT-3 with

Symptom	Contribution	Symptom	Contribution
Anhedonia	0.13	Psychomotor	0.10
Concentration	0.05	Self-esteem	0.09
Eating	0.05	Self-harm	0.17
Fatigue	0.13	Sleep	0.04
Mood	0.09		

Table 6: Contribution of symptoms to depression detection.

prompted examples. We prompt GPT-3 with four examples for each (*positive*, *negative*) class from one dataset (e.g., TRT) and evaluate on other datasets (e.g., RSDD, eRisk). Due to the high computational cost of GPT-3, we only evaluate on 100 positive samples and 100 negative samples from each dataset.

We can see in Table 7 that prompted GPT-3 is consistently outperformed by our classifier-constrained methods, and the margin is often large. For example, among models trained on RSDD, the classifier-constrained model with CNN vectors achieves the highest F1 of 0.79 and 0.64 when tested on TRT and eRisk, respectively. GPT-3 performs worse with at least a 0.12 drop in F1. This result demonstrates that depression detection is still challenging for large few-shot learners, further highlighting our contributions of generalizable methods. However, we note that this setting has several limitations that prevent a completely fair comparison. Our methods have access to hundreds of posts, while GPT-3 has a limitation on the prompt length. In addition, prompt examples, which have high influence on the GPT-3 few-shot performance, need to be carefully selected and tuned. It is possible that we were unable to identify near-optimal prompts. Furthermore, it is difficult to know which posts or users should be prompted to GPT-3, so we opted to select randomly. Lifting this limitation would require a separate model to identify which posts should be used as input.

5 Case study

In Table 3, we demonstrate approaches trained on TRT using text from an anonymized and paraphrased depressed user from the eRisk2018 dataset. We show the top two posts ranked by the drop in depression score when excluding each post. All models were able to produce correct labels with very high confidence. However, there is a clear difference in the posts that models rely primarily on for prediction. The PHQ9 (scores) model found highly relevant posts with convincing associated symptoms. For example, in the first post, the PHQ9 models found 3 symptoms, including "*anhedonia*", "*mood*" and "*self-esteem*". By looking at those posts and symptoms, mental health professionals could quickly understand the patient's circumstances and make further decisions. The two most important posts for PHQ9Plus and BERT are more about daily life concerns or complaints, which may be less useful to explain a high depression score than the top posts used to explain the PHQ9 (scores) model. While these posts are relevant, they are more difficult to interpret than posts directly mentioning symptoms that are known to be relevant. Furthermore, in the TRT training dataset, due to the biased selection of control users, those life concerns/complaints may form a shortcut that effectively differentiates depressed users from control users. However, in more realistic deployment scenarios (i.e. *dataset-transfer* settings), the fact that such shortcuts do not generalize makes PHQ9Plus and BERT more unreliable and fragile.

Prompt/Train RSDD	Test TRT	Test eRisk
PHQ9 (scores)	0.64	0.62
PHQ9 (vectors)	**0.79**	**0.64**
GPT-3	0.59	0.52
Prompt/Train TRT	**Test RSDD**	**Test eRisk**
PHQ9 (scores)	**0.71**	0.72
PHQ9 (vectors)	0.69	**0.78**
GPT-3	0.61	0.54
Prompt/Train eRisk	**Test RSDD**	**Test TRT**
PHQ9 (scores)	**0.85**	**0.74**
PHQ9 (vectors)	0.83	0.71
GPT-3	0.54	0.49

Table 7: F1 scores of PHQ9 models vs. GPT-3

6 Conclusion

In this work, we propose a spectrum of methods for depression detection that are constrained by the presence of PHQ9 symptoms. In our experiments on the three datasets, we find these methods to perform well compared to strong baselines while generalizing better to similar datasets. This can be viewed as a proof-of-concept demonstrating that grounding depression predictions in PHQ9 can improve the generalizability of depression detection and the interpretability of the model. While this research focuses only on depression detection, the idea of constraining models to consider only relevant causes may be applied to a wider range of tasks, including detection of other mental health conditions with diagnostic questionnaires.

Ethics Statement

Due to the sensitivity of the mental health related data, additional consideration needs to be taken into account when accessing and analyzing such data, as highlighted by Benton et al. (2017). All datasets used in this research were obtained according to each dataset's respective data usage policy. We did not interact with users in any way, and we refrained from showing any direct excerpts of the data in this manuscript to prevent risks from identifying users' pseudonyms. (All excerpts have been paraphrased.) Similarly, we made no attempt to identify, deanonymize, or link users to other social media accounts. These precautions ensure we do not draw attention to specific users who may be suffering from depression.

All models proposed in this research were trained on social media data. Thus, they are likely to fail on data coming from other sources (e.g., clinical notes), and there are no accuracy guarantees even within social media data. Our models are not intended to replace clinicians. Instead, we envision the approaches we describe being used as assistive tools by mental health professionals.

References

Hessam Amini and Leila Kosseim. 2020. Towards explainability in using deep learning for the detection of anorexia in social media. In *Natural Language Processing and Information Systems*, pages 225–235, Cham. Springer International Publishing.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 94–102, Valencia, Spain. Association for Computational Linguistics.

Louise Brådvik. 2018. Suicide risk and mental disorders.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Louise Brådvik. 2018. Suicide risk and mental disorders.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 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*, pages 1485–1497.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60, Baltimore, Maryland, USA. Association for Computational Linguistics.

Glen Coppersmith, Ryan Leary, Patrick Crutchley, and A. Fine. 2018. Natural language processing of social media as screening for suicide risk. *Biomedical Informatics Insights*, 10.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. 2020. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. *ICWSM*, 13:1–10.

Fionn Delahunty, Robert Johansson, and Mihael Arcan. 2019. Passive diagnosis incorporating the PHQ-4 for depression and anxiety. In *Proceedings of the Fourth Social Media Mining for Health Applications (#SMM4H) Workshop & Shared Task*, pages 40–46, Florence, Italy. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. 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)*, pages 4171–4186.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. 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), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sindhu Kiranmai Ernala, Michael L. Birnbaum, Kristin A. Candan, Asra F. Rizvi, William A. Sterling, John M. Kane, and Munmun De Choudhury. 2019. 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*, CHI '19, page 1–16, New York, NY, USA. Association for Computing Machinery.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson H S Liu, Matthew E. Peters, Michael Schmitz, and Luke Zettlemoyer. 2017. A deep semantic natural language processing platform.

Robert Geirhos, Jorn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009.

Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. 2014. The distress analysis interview corpus of human and computer interviews. In *LREC*, pages 3123–3128.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2020. Do models of mental health based on social media data generalize? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3774–3788, Online. Association for Computational Linguistics.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2021. On the state of social media data for mental health research. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 15–24, Online. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanisław Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp.

Zheng Ping Jiang, Sarah Ita Levitan, Jonathan Zomick, and Julia Hirschberg. 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*, pages 147–156.

Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The PHQ-9: validity of a brief depression severity measure. *Journal of general internal medicine*, 16(9):606–613.

Andrew Lee, Jonathan K. Kummerfeld, Larry An, and Rada Mihalcea. 2021. Micromodels for efficient, explainable, and reusable systems: A case study on mental health. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4257–4272, Punta Cana, Dominican Republic. Association for Computational Linguistics.

David E Losada and Fabio Crestani. 2016. A test collection for research on depression and language use. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 28–39. Springer.

David E Losada, Fabio Crestani, and Javier Parapar. 2019. Overview of erisk 2019 early risk prediction on the internet. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 340–357. Springer.

Matthew Matero, Akash Idnani, Youngseo Son, Salvatore Giorgi, Huy Vu, Mohammad Zamani, Parth Limbachiya, Sharath Chandra Guntuku, and H Andrew Schwartz. 2019. Suicide risk assessment with multi-level dual-context language and bert. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 39–44.

David N. Milne, Glen Pink, Ben Hachey, and Rafael A. Calvo. 2016. CLPsych 2016 shared task: Triaging content in online peer-support forums. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 118–127, San Diego, CA, USA. Association for Computational Linguistics.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana. Association for Computational Linguistics.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. *BioNLP 2019*, page 58.

Nairan Ramirez-Esparza, Cindy K Chung, Ewa Kacewicz, and James W Pennebaker. 2008. The psychology of word use in depression forums in english and in spanish: Texting two text analytic approaches. In *ICWSM*.

Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2020. Adapt or get left behind: Domain adaptation through bert language model finetuning

for aspect-target sentiment classification. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4933–4941.

Alex Rinaldi, Jean E Fox Tree, and Snigdha Chaturvedi. 2020. Predicting depression in screening interviews from latent categorization of interview prompts. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7–18.

Han-Chin Shing, Suraj Nair, Ayah Zirikly, Meir Friedenberg, Hal Daumé III, and Philip Resnik. 2018. Expert, crowdsourced, and machine assessment of suicide risk via online postings. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 25–36, New Orleans, LA. Association for Computational Linguistics.

Han-Chin Shing, Philip Resnik, and Douglas W Oard. 2020. A prioritization model for suicidality risk assessment. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8124–8137.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Jt Wolohan, Misato Hiraga, Atreyee Mukherjee, Z. Sayyed, and Matthew Millard. 2018. Detecting linguistic traces of depression in topic-restricted text: Attending to self-stigmatized depression with nlp.

Shweta Yadav, Jainish Chauhan, Joy Prakash Sain, Krishnaprasad Thirunarayan, Amit Sheth, and Jeremiah Schumm. 2020. Identifying depressive symptoms from tweets: Figurative language enabled multitask learning framework. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 696–709, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978, Copenhagen, Denmark. Association for Computational Linguistics.

Ayah Zirikly, Philip Resnik, Ozlem Uzuner, and Kristy Hollingshead. 2019. CLPsych 2019 shared task: Predicting the degree of suicide risk in Reddit posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 24–33.

A Details of experimental setup

We designed various experiments to validate and analyze two main components: the questionnaire and depression models. The hyper-parameters of those models were set as follows:

Questionnaire model: In classifier-constrained methods, we trained a CNN for each of the nine symptom classifiers. We used filters of sizes $[2, 3, 4, 5, 6]$, and one filter for each size. Consequently, the max-pooling produces a vector of size 5, which is then fed into the final linear layer for prediction. We apply L_2 regularization specifically to the CNN kernels. The L_2 weights were fine-tuned with three options $[0.1, 0.01, 0.001]$.

Depression model: We experimented with 6 variations described in Table 1. All of these variations, except for the first one, use a CNN classifier on top of different inputs ranging from BERT embedding to pattern scores. The CNN used here has $(filter\_size = [2, 3, 4, 5, 6], num\_filter = 50)$, and $(k = 5)$ for k-max pooling. The threshold for the first variation was tuned on from 1 to 10.

In all experiments with pre-trained BERT, we used the BERT-base version (Devlin et al., 2019a). We do not fine-tune BERT's parameters since in a pilot study, we found that fine-tuning BERT does not improve the generalization. All models were optimized using a cross-entropy loss with class weights of 0.1 and 0.9 for the control and depressed classes, respectively, a 1cycle learning rate scheduler (with a maximum of 0.01), batches of size 64, and early stopping after five epochs. When calculating F1, we set the decision threshold to 0.5, because it cannot be safely tuned in dataset-transfer experiments.

B Manually-labeled questionnaire dataset

To evaluate the performance of our weakly-supervised symptom classifiers, we prepared 900 examples manually labeled by three annotators. The labeled samples were randomly selected posts containing carefully selected keywords (e.g., keywords that are close to positive patterns - "*sleep*" in "*can't sleep*") to avoid including too many easy true negatives. The labeling process involved three annotators. The first annotator labeled all 100 instances, and the second annotator re-labeled 50 of them. If the agreement on twice-labeled examples was weak ($\kappa < 0.60$), the second annotator would continue to annotate the remaining 50 examples. The third annotator adjudicated label dis-

Symptom	# patterns	Symptom	# patterns
Anhedonia	116	Psychomotor	108
Concentration	70	Self-esteem	102
Eating	145	Self-harm	126
Fatigue	73	Sleep	118
Mood	110		

Table 8: Number of patterns for each question

agreements between the first two annotators.

C Details of questionnaire dataset construction

This section describes the data creation process for the questionnaire models, which consists of 9 symptom classifiers corresponding to 9 questions in the PHQ9 instrument (e.g., *"trouble falling or staying asleep?"*). PHQ9 questions ask how often the patient experienced each symptom; we adapt this approach to our domain by classifying whether a given post contains a symptom. Given the lack of training data for this task, we collected regular expression patterns and used these patterns together with heuristics to construct training data (positive class and negative class) for each symptom.

C.1 Positive class

For each question in the PHQ9 diagonstic instrument, we prepared a set of positive patterns that each indicates the presence of a symptom described by the question. For example, the patterns "*don'?t feel like doing anything*" and "*can'?t fall back to sleep*" describe the anhedonia and sleep symptoms, respectively. The number of patterns for each symptom is shown in Table 8. Each pattern set is then matched against a collection of posts crawled from 127 mental-health subreddits[3] and also against posts from the SMHD dataset (Cohan et al., 2018), which was constructed from Reddit but excludes mental health subreddits. The purpose of using these two raw datasets is to increase the diversity and minimize the bias of the data. In the labeling step, if a post contains a match with any positive pattern of a question, we select that post as a positive training sample for the corresponding symptom question.

While pattern matching is fast and transparent, it is inflexible and may produce many false positives (FP). Below, we introduce the four most popular FP cases discovered in our analysis and the techniques we employed to mitigate them when constructing weakly-labeled training data.

[3] https://files.pushshift.io/reddit/

Positive sentiment. Some posts contain positive patterns but do not show depressive signal; for example, "*Friends and I stayed up all night playing a game*" contains the positive pattern "*stayed up all night*", but it shows excitement about the game rather than a sleep issue. As a solution, we removed all posts containing positive sentiment with the help of Allen NLP's sentiment model (Gardner et al., 2017).

Conditional clause. Sometimes users hypothesize about their health conditions, such as "*If I lost my appetite for days at a time, that... wouldn't be sustainable for me.*". We remove these posts by using regular expressions to identify popular conditional clause formats.

Third-person pronouns. Users may attribute a condition to someone else, such as in "*he is easily distracted.*" We identified posts of this kind by checking whether the closest pronoun to the positive pattern is third-person or first-person, and removing posts in the former category.

Negation. Positive patterns may be negated, such as in "*haven't had a suicidal thought in ages.*". To handle this situation, we removed posts containing positive patterns preceded by a negation word, such as ("*not*", "*never*", '*rarely*", etc.).

C.2 Negative class

Identifying hard negative samples is crucial for the quality of the trained classifiers. Models trained on negative examples that are too easy might be prone to over-fitting or perform only keyword matching. Therefore, we propose five heuristics for identifying and synthesizing negative examples.

Keywords. We collect negative posts that contain some keywords, such as *"sleep"*, but do not contain a positive pattern ("*can't sleep*"). This hinders models to perform simple keyword matching.

Pronouns. We replace the first-person pronouns appearing in posts from the positive class with third-person pronouns or proper nouns, such as replacing "*I*" with "*she*."

Other symptoms. We use randomly selected posts labeled positive for other questions (symptoms) as negative examples for the given question. We ensure the selected posts do not contain positive patterns for the given question.

Negation. For each positive pattern defined in the previous section, we created a corresponding negated one, such as negating "*have sleep apnea*" to "*never have sleep apnea*". Only matched sen-

tences were used in this method, because using the whole post with only some sentences being negated could lead to contextual inconsistencies.

Positive sentiment. We use training instances labeled neutral or positive in a sentiment dataset as negative examples. In particular, we used the Sentiment140 corpus, which contains 1.6 millions tweets (Go et al., 2009).