A Simple and Flexible Modeling for Mental Disorder Detection by Learning from Clinical Questionnaires

Hoyun Song **Jisu Shin** **Huije Lee** **Jong C. Park**[*]
School of Computing
Korea Advanced Institute of Science and Technology
{hysong1991,jisu.shin,angiquer,jongpark}@kaist.ac.kr

Abstract

Social media is one of the most highly sought resources for analyzing characteristics of the language by its users. In particular, many researchers utilized various linguistic features of mental health problems from social media. However, existing approaches to detecting mental disorders face critical challenges, such as the scarcity of high-quality data or the trade-off between addressing the complexity of models and presenting interpretable results grounded in expert domain knowledge. To address these challenges, we design a simple but flexible model that preserves domain-based interpretability. We propose a novel approach that captures the semantic meanings directly from the text and compares them to symptom-related descriptions. Experimental results demonstrate that our model outperforms relevant baselines on various mental disorder detection tasks. Our detailed analysis shows that the proposed model is effective at leveraging domain knowledge, transferable to other mental disorders, and providing interpretable detection results.

1 Introduction

Mental health problems, a significant challenge in public healthcare, are usually accompanied by distinct symptoms, such as loss of interest or appetite, depressed moods, or excessive anxiety. As these symptoms can often be expressed over social media, detecting mental health conditions using social media text has been studied extensively (Yates et al., 2017; Coppersmith et al., 2018; Matero et al., 2019; Murarka et al., 2021; Harrigian et al., 2021; Jiang et al., 2021; Nguyen et al., 2022). Such approaches could give rise to a monitoring system that provides clinical experts with information about possible mental crises.

To automatically identify mental health problems, traditional approaches focus on finding linguistic patterns and styles from the language of psychiatric patients. Utilizing these features, statistical models can explain the correlation between linguistic factors and mental illnesses. However, these approaches suffer from increased complexity of models, necessitating pipelines of steps, from engineering features to producing results. By contrast, more recent works have employed strong pretrained models, which allow a direct use of raw data and simplify model development (Matero et al., 2019; Jiang et al., 2020). While such end-to-end approaches may be effective at achieving higher performance, they often lack domain-based interpretation, which is essential for decision-support systems (Mullenbach et al., 2018). Hence, there is a trade-off between providing interpretable predictions based on domain knowledge and the simplicity of the models.

The lack of a sufficient sample size for high-quality data is another challenge in the clinical domain (De Choudhury et al., 2017; Harrigian et al., 2020). Despite the availability of diverse datasets and methods for detecting mental disorders, most of them aim primarily at identifying only clinical depression. To tackle such a problem, recent studies have focused on developing transferable linguistic features that can be used for the detection of various mental disorders (Aich and Parde, 2022; Uban et al., 2022). However, the linguistic features that are trained on a particular dataset may not be fully transferable to a different task (Ernala et al., 2019; Harrigian et al., 2020).

Others utilized symptom-related features that are more common properties of psychiatric patients, resulting in generalizability of depression detection (Nguyen et al., 2022). Despite this improvement, however, their approach still faces challenges because they rely on pipelined methods using manually-defined symptom patterns. Such symptom patterns for depression detection lack flexibility as they cannot be easily adapted to other mental disorders. In addition, the pipeline approach

[*] Corresponding author

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 12190–12206
July 9-14, 2023 ©2023 Association for Computational Linguistics

with symptom extraction is quite complex to implement. It involves multiple steps, designing symptom patterns, training a symptom identification model, and detecting depression using the identified symptom patterns.

To address these challenges, we propose to design a simple and more flexible approach that also preserves interpretability. We are motivated by the process that humans use to quickly learn related features, often by reading just a single explanation. For example, when people are reading depression questionnaires, they readily understand the questions and learn about symptoms that are related to depression, allowing them to self-diagnose their levels of depression.

To this end, we employ the siamese network (Koch et al., 2015), which captures the semantic meaning of the text inputs and compares them directly to symptom-related descriptions. This process is simple since they find symptom-related clues directly from the input, rather than relying on hand-engineered features or intermediate models. Our proposed model, Multi-Head Siamese network (MHS), can be easily adapted to other mental illness domains by simply replacing the symptom-related descriptions. In addition, our model is designed to capture the distinct features of each symptom using multiple heads. By examining the learned weights of each symptom head, our model gives rise to human-understandable interpretations.

We evaluate the performance of our model, detecting texts containing mental health problems on four mental disorders. Furthermore, the detailed analysis of the proposed model shows its efficiency in utilizing symptom-related knowledge, its ability to be applied to different mental disorders, and its interpretable reasoning for detected results.

2 Related Work

Social media are commonly used for mental health research because of the ease of access to various aspects of human behavior studies. Similarly to other NLP domains, pre-trained language models, such as BERT (Devlin et al., 2019), are widely used for identifying mental health problems (Matero et al., 2019; Jiang et al., 2020; Murarka et al., 2021; Dinu and Moldovan, 2021).

Others have presented interpretable detection methods for the mental health domain based on linguistic features (Song et al., 2018; Uban et al., 2021). Various efforts have also been made to study such linguistic features accompanying mental illness, such as differences in word usage (Tadesse et al., 2019; Jiang et al., 2020; Dinu and Moldovan, 2021), or in syntactic features (Kayi et al., 2017; Ireland and Iserman, 2018; Yang et al., 2020). Some studies address the differences between sentiments or emotional aspects (Preoţiuc-Pietro et al., 2015; Kirinde Gamaarachchige and Inkpen, 2019; Allen et al., 2019; Wang et al., 2021), or differences in topics (Tadesse et al., 2019; Kulkarni et al., 2021).

The linguistic features are also used for transferable methods across other mental disorders (Aich and Parde, 2022; Uban et al., 2022), focusing on the fact that a large number of studies have been done primarily on depression (De Choudhury et al., 2013; Yates et al., 2017; Eichstaedt et al., 2018; Song et al., 2018; Tadesse et al., 2019; Yang et al., 2020; Nguyen et al., 2022), compared to other disorders, such as anxiety disorder (Ireland and Iserman, 2018), anorexia (Uban et al., 2021), or schizophrenia (Kayi et al., 2017). However, such linguistic features do not generalize well to new user groups. For example, De Choudhury et al. (2017), Loveys et al. (2018), and Pendse et al. (2019) found that the linguistic styles may vary to their backgrounds. In addition, Harrigian et al. (2020) found that a model trained on a particular dataset does not always generalize to others. To handle such a generalization problem, Nguyen et al. (2022) and Zhang et al. (2022) focused on the shared and general properties (i.e., symptoms) of a mental health problem. However, unlike ours, which captures the symptom features directly from raw data, these methods require additional steps for learning symptom-related features.

In this paper, we use the siamese network (Koch et al., 2015), based on one-shot learning, exploited recently for simple networks (Chen and He, 2021; Zhu et al., 2021). We utilize the symptom descriptions sourced from DSM-5 (American Psychiatric Association, 2013) to make our model learn symptom-related knowledge.

3 Methodology

In this section, we introduce our simple but flexible modeling for leveraging clinical questionnaires. Our model aims to detect texts with mental illness episodes based on the presence of symptom-related features just by a single component.

An overview of our network is shown in Figure 1.

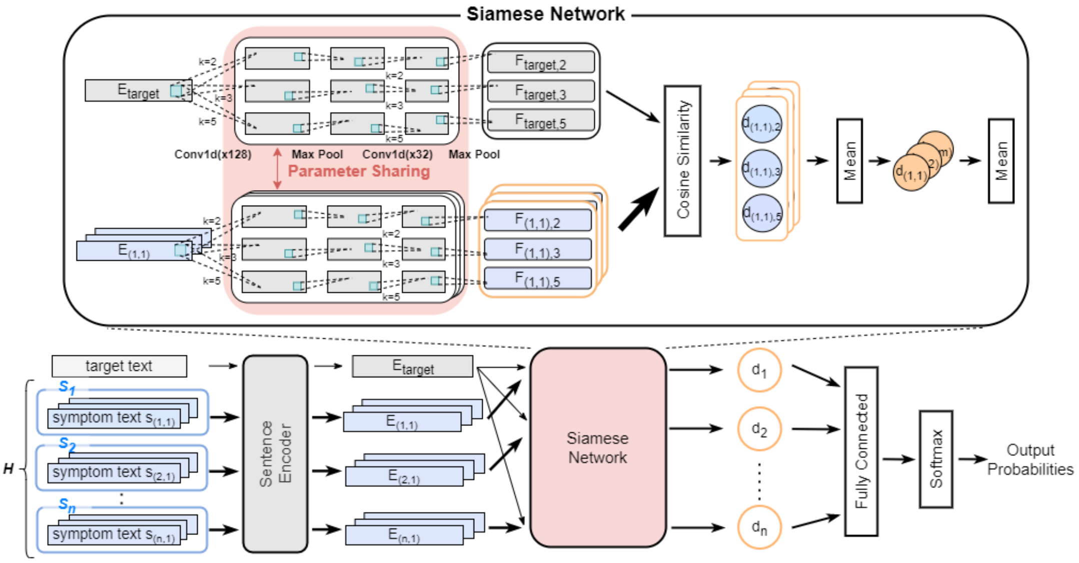

Figure 1: The model architecture of Multi-Head Siamese network (MHS). S_i indicates a head of a symptom that contains the symptom-related descriptions $s_{(i,j)}$, and d_i indicates a distance value computed by cosine similarity between the target text and the descriptions. MHS compares the contextualized embeddings of the target text and symptom and predicts the probability of mental illness based on context similarity.

We designed our model based on the siamese network (Koch et al., 2015). As with the original siamese neural network, our model also contains a single feature extractor with shared parameters. The extractor directly obtains features from contextualized embeddings generated by sentence encoders. Then, employing the similarity function, we compare the similarity to see the presence of symptom-related features from the target text. In addition, we apply multi-headed learning to the original siamese network, repeating the comparison process for each distinct symptom. We describe the detailed structure in the following subsections.

3.1 Model Structure

Our model, the Multi-Head Siamese network (MHS), is an end-to-end model that takes raw input texts and produces the final result without the need for manual feature engineering. MHS is designed to take two types of inputs, the target text to be classified and descriptions of symptoms. The descriptions are grouped for each symptom, and each symptom group is the input for the corresponding symptom head. For example, assuming that we have n symptoms for discriminating against mental disorder, we build a set of n heads (H) from S_1 to S_n for the detection model as follows:

$$H = \{S_1, S_2, ..., S_n\} \tag{1}$$

Each head S represents discrete symptoms, containing a number of descriptions and questions regarding the corresponding symptom. For example, if S_i has m sentences describing the symptom, we have a set S_i of questions:

$$S_i = \{s_{(i,1)}, s_{(i,2)}, ..., s_{(i,m)}\} \tag{2}$$

With a given input of the target sentence, our model obtains embedding vectors (E_{target}) by employing pre-trained sentence encoders, such as BERT or RoBERTa. We also get symptom embeddings by encoding all sentences from all heads (H).

Our siamese network employs a multi-channel convolutional neural network (CNN) for feature learning. We apply three channels for convolution layers, whose kernel sizes are 2, 3, and 5. Thus, our model is designed to capture informative clues with the window sizes of 2, 3, and 5 from texts. Each channel contains two convolutional layers and two max-pooling layers. The final convolutional layer is flattened into a single embedding vector. As a result, we obtain three feature embedding vectors ($F_{target,k}$) with $k = 2, 3, 5$ from the target text:

$$F_{target,k} = Conv1d_k(E_{target}) \tag{3}$$

Through the same process, we also obtain feature embedding vectors from symptom texts from the i^{th} head and j^{th} sentence as follows:

$$F_{(i,j),k} = Conv1d_k(E_{(i,j)}) \tag{4}$$

We compute the distances (d) between the target feature vector ($F_{target,k}$) and a symptom-sentence vector ($F_{(i,j),k}$) using cosine similarity, ranging from $[-1, 1]$. We calculate a single distance value by taking the average of K distance values, where K represents the number of channels:

$$sim(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{xy}}{\|\mathbf{x}\| \|\mathbf{y}\|} \tag{5}$$

$$d_{(i,j)} = \frac{1}{K} \sum_{k} sim(F_{target,k}, F_{(i,j),k}) \tag{6}$$

Finally, when there are distance values for all sentences, they are averaged to yield the distance value of the i^{th} head (d_i):

$$d_i = \frac{1}{m} \sum_{j=1}^{m} d_{(i,j)} \tag{7}$$

To regularize the results, we choose to use averaging as an aggregation function for the distance values.

We iterate this process over the number of heads (n). After the siamese network step, all distance values (d_i) are stacked into a $1 \times n$ vector (D). By applying the fully connected layer, the distance vector is reduced into a two-dimensional vector o, which is an output probability of classifying mental illness:

$$f : \mathbb{R}^n \rightarrow \mathbb{R}^2 \tag{8}$$

$$o = f(D) = W^T \cdot D + b \tag{9}$$

By analyzing the weights (W) and distance values (D) of the fully connected layer, we can examine which symptoms are activated as important information when classifying the related mental disorder. Further details are discussed in Section 5.4. The implementation code and symptom-sentences are made publicly available[1].

3.2 Symptom Descriptions

In the present study, we focus on four mental disorders: major depressive disorder (MDD), bipolar disorder, generalized anxiety disorder (GAD), and borderline personality disorder (BPD). As summarized in Table 1, we compiled the diagnostic criteria for each mental disorder, sourced from DSM-5. We constructed heads based on the list of symptoms. For example, in the case of MDD, there are a total of 9 symptoms (D0-D8), so when constructing a model detecting depressive symptoms, there will be a total of 9 heads ($n(H_{MDD}) = 9$). As for bipolar disorder, symptoms can be divided into depressive episodes (D0-D8) and manic episodes (M0-M7), with a total of 17 heads. The depressive episodes of bipolar disorder are the same as those of MDD.

Mental Disorders	**Diagnostic Criteria from DSM-5**
Major Depressive Disorder (D0-D8)	D0. Depressed mood most of the day
	D1. Diminished interest or pleasure
	D2. Sleep disorders (insomnia or hypersomnia)
	D3. Changes in weight or appetite when not dieting
	D4. Fatigue or loss of energy
	D5. Feeling worthlessness or guilty
	D6. Diminished ability to think or concentrate
	D7. A slowing down of thought and a reduction of physical movement
	D8. Recurrent thoughts of death and suicidal ideation
Bipolar Disorder (D0-D8, M0-M7)	**Major Depressive Episode**
	D0-D8: Same as major depressive disorder
	Manic Episode
	M0. A distinct period of persistently elevated or expansive mood
	M1. Increase in goal-directed activity
	M2. Inflated self-esteem or grandiosity
	M3. Decreased need for sleep
	M4. More talkative than usual
	M5. Flight of ideas
	M6. Distractibility
	M7. Activities that have a high potential for painful consequences
Generalized Anxiety Disorder (A0-A6)	A0. Excessive anxiety and worry more than 6 months
	A1. Difficult to control the worry
	The anxiety and worry are associated with followings:
	A2. Irritability
	A3. Being easily fatigued
	A4. Sleep disturbance
	A5. Difficulty concentrating or mind going blank
	A6. Muscle tension
Borderline Personality Disorder (B0-B8)	B0. Interpersonal relationships alternating between idealization and devaluation
	B1. Recurrent suicidal or self-mutilating behavior
	B2. Identity disturbance
	B3. Affective instability
	B4. Inappropriate anger or difficulty controlling anger
	B5. Transient, stress-related paranoid ideation or severe dissociative symptom.
	B6. Impulsive behaviors that are potentially self-damaging
	B7. Frantic efforts to avoid abandonment
	B8. Chronic feelings of emptiness

Table 1: A summary of diagnostic criteria for each mental disorder, sourced from DSM-5.

Each head includes a description of diagnostic criteria and questions from self-tests corresponding to each symptom. As a result, each head contains two or more sentences ($n(S) \geq 2$). In the case of more than two related questions for a symptom, the corresponding head contains more than two sentences.

We collected the questions from the publicly available self-tests[2]. The process was conducted under the guidance of a psychology researcher. The complete list of collected sentences for each head is shown in Appendix C. Our model can easily

[1] https://github.com/HoyunSong/acl23-multi-head-siamese-mental-illness

[2] MDD (www.psycom.net/depression-test/),
Bipolar (www.psycom.net/bipolar-disorder-symptoms/),
GAD (www.psycom.net/anxiety-test), and
BPD (www.psycom.net/borderline-personality-test/)

Subreddit	**#samples**	**sent.**	**tok.**	**vocab.**
r/depression	11,416	9.5	143.1	43,766
r/bipolar	10,941	10.5	157.1	54,426
r/anxiety	11,471	9.7	159.8	51,936
r/bpd	10,979	11.8	187.5	53,741
Random	40,570	8.8	123.0	198,988
Total	85,377	9.6	133.6	229,309

Table 2: The number of samples, average numbers of sentences and tokens, and the vocabulary size.

transfer to other mental disorders by just replacing symptom descriptions, as evidenced by the findings in Section 5.3.

4 Experiments

4.1 Dataset and Evaluation

In order to evaluate our model, we constructed four datasets to detect possible mental disorder episodes. We sampled posts from Reddit[3], which is one of the largest online communities. Each sample is a concatenation of a title and a body from a post. Each dataset contains two groups of Reddit posts. One includes the posts collected from mental disorder-related subreddits as a text containing the mental illness contents, and the other is from random subreddits as a clean text. The detailed statistics of each group is shown in Table 2. We performed preprocessing by discarding posts containing URLs or individually identifiable information, and posts shorter than ten words (i.e., tokens). We only retained posts in English; otherwise, they are discarded.

We conducted four tasks, employing these collected datasets, discriminating texts sourced from mental disorder-related subreddits out of non-mental illness texts. The details of each task are as follows: MDD detection (*r/depression*+random), Bipolar disorder detection (*r/bipolar*+random), GAD detection (*r/anxiety*+random), and BPD detection (*r/bpd*+random).

To compare our model with baseline models with respect to classification performance, we report results using standard metrics, Accuracy (Acc.), F1 score (F1) for the mental illness group, and Area Under the Curve (AUC). The performance measure is reported by five-fold cross-validation, and each repetition is trained on six different seeds. We averaged after 30 runs (5×6) to get the final result.

4.2 Baselines and Experimental Setup

In this subsection, we describe models and implementation details for experiments. More experimental details are shown in Appendix A.

1) Traditional Models We implemented two feature-based classifiers, a support vector machine (SVM) and a random forest (RF), with two versions: **BoW**, employing lexical features only (Tadesse et al., 2019; Jiang et al., 2020), and **Feature**, adding sentimental and syntactic features (Allen et al., 2019; Yang et al., 2020; Wang et al., 2021). **2) BERT** (Devlin et al., 2019) is one of the most well-known baseline models using contextualized embeddings (Jiang et al., 2020; Matero et al., 2019). **3) XLNet** (Yang et al., 2019) is another strong baseline with a pre-trained language model (Dinu and Moldovan, 2021). **4) RoBERTa** (Liu et al., 2019) is a robustly optimized BERT and one of the most solid baselines in natural language classification (Dinu and Moldovan, 2021; Murarka et al., 2021). **5) GPT-2** (Radford et al., 2019) is a strong few-shot learner with a large Transformer-based language model. **6) PHQ9** (Nguyen et al., 2022) is a depression detection model constrained by the presence of PHQ9 symptoms.

We implemented our models using PyTorch and fine-tuned our models on one 24GB Nvidia RTX-3090 GPU, taking about 13 minutes for each epoch. The batch size and embedding size of all models are 8 and 512, respectively, and are fine-tuned over five epochs. We truncated each post at 512 tokens for all models. For each model, we manually fine-tuned the learning rates, choosing one out of {1e-5, 2e-5, 1e-6, 2e-6} that shows the best F1 score. We report the average results over 30 runs (five-fold cross-validations are trained on six different seeds) for the same pre-trained checkpoint.

4.3 Experimental Results

Table 3 shows the overall performance of our proposed model (MHS) and strong baselines on four tasks. Each task is about detecting texts with corresponding mental illness episodes on social media. We see that our model outperforms all competing approaches, including linguistic feature-based models, end-to-end pre-trained models, and a method that uses symptom-related knowledge.

Linguistic feature-based models exhibit significant performance variations based on the level of detail in their feature design. By contrast, MHS can simply find the features directly from the contextualized representation, giving better performance improvements. Pre-trained models with contextualized embeddings have the benefits that can be

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

Model	MDD			Bipolar			GAD			BPD		
	Acc.	F1 (±)	AUC	Acc.	F1 (±)	AUC	Acc.	F1 (±)	AUC	Acc.	F1 (±)	AUC
RF-BoW	89.9	73.7 (0.34)	80.4	90.9	75.8 (0.37)	81.1	91.7	76.3 (0.41)	81.7	90.3	73.2 (0.35)	79.8
SVM-Bow	91.2	78.0 (0.89)	83.6	90.2	78.2 (0.84)	81.4	92.9	83.3 (0.84)	88.5	93.4	83.6 (0.67)	88.9
RF-Feature	89.6	72.9 (0.54)	79.8	91.1	76.2 (0.54)	81.4	91.8	79.2 (0.77)	83.7	90.4	73.5 (0.45)	80.0
SVM-Feature	92.2	81.5 (0.59)	86.6	93.3	83.6 (0.77)	87.5	94.3	86.7 (0.81)	90.0	93.6	83.6 (0.41)	88.6
GPT-2	94.6	88.0 (0.51)	92.6	95.3	88.9 (0.63)	92.4	95.7	90.2 (0.35)	93.5	95.6	89.7 (0.49)	93.4
XLNet	94.4	87.9 (0.40)	92.1	95.2	88.8 (0.43)	92.4	95.7	89.8 (0.26)	93.2	95.6	89.4 (0.43)	92.9
BERT	94.2	87.3 (0.41)	92.4	95.0	88.1 (0.56)	91.3	95.3	88.5 (0.61)	91.9	95.0	88.9 (0.55)	93.2
BERT-PHQ9	94.4	87.2 (0.47)	91.8	95.2	88.4 (0.48)	91.8	95.2	88.2 (0.48)	91.4	95.1	88.9 (0.46)	92.5
BERT-MHS	94.9	88.6 (0.29)	93.0	95.4	89.2 (0.42)	92.3	95.7	90.3 (0.38)	93.7	95.7	90.0 (0.28)	93.7
RoBERTa	94.8	88.6 (0.34)	93.1	95.4	89.4 (0.56)	92.9	95.8	90.4 (0.35)	93.7	95.7	90.3 (0.35)	93.7
RoBERTa-PHQ9	94.9	88.6 (0.50)	92.6	95.4	89.4 (0.59)	92.6	95.5	89.4 (0.33)	92.4	95.6	89.9 (0.47)	93.3
RoBERTa-MHS	**95.5**	**89.6 (0.31)***	**93.8**	**95.8**	**90.4 (0.31)***	**93.4**	**96.2**	**91.5 (0.28)***	**94.3**	**95.9**	**90.8 (0.26)***	**94.0**

Table 3: Results on four mental disorder detection tasks. Each result is averaged after 30 runs. The best results for each task are shown in bold, and the second-best results are underlined. * denotes that the performance gain is statistically significant with $p < 0.05$ under all pairwise t-tests.

Model	**#parameters**	**Relative Size**
BERT	108,311,810	1.00
MHS w/bert	108,967,319	1.01
RoBERTa	124,647,170	1.15
MHS w/roberta	125,302,679	1.16

Table 4: The numbers of parameters for BERT, RoBERTa, and our models.

Model	Acc.	Pre.	Rec.	F1	AUC
CNNs w/bert emb.	94.0	**89.8**	82.9	86.2	90.1
+single-head	94.5	88.6	86.8	87.6	91.7
+multi-head +one description	94.9	87.3	90.2	88.7	93.2
+multi-head +multi-description	**95.4**	89.1	**90.5**	**89.7**	**93.9**

Table 5: An ablation study of different levels of knowledge and features affecting our model. The result is the average of the four tasks.

easily fine-tuned for a wide range of tasks. However, compared to MHS, they lack a specific focus on domain-based features, while MHS is tailored to identify such features, leading to better performance.

We implemented our model and PHQ9 model with two different encoders, BERT and RoBERTa, and the tendency for performance improvement is the same on both encoders. Both PHQ9 and MHS leverage symptom-related information but differ in their architecture, specifically whether it is a multi-step pipeline or an end-to-end model. The end-to-end design of MHS allows for direct learning of complex relationships, reducing the potential for error propagation, and resulting in enhanced performance compared to the pipeline model. Moreover, for this pipeline model to apply to other mental disorders, a symptom pattern must be created for each mental disorder, which is challenging to achieve without expert-level knowledge. On the other hand, our proposed model overcomes these challenges by simply replacing symptom descriptions. A detailed analysis of the performance improvement is shown in Section 5.

4.4 Model Parameters

Table 4 shows the number of parameters for each model. Compared to the baseline models, the additional number of parameters for our siamese network is about 655K. It is a much smaller number than that of the additional parameters for RoBERTa and BERT (about 16M), but the performance of MHS (w/bert) is slightly better or shows little difference. It suggests that our proposed model, learning domain knowledge, achieves much efficient performance improvement by adding just a small number of parameters.

5 Model Analysis and Discussions

5.1 Ablation Study

We conducted an ablation study to investigate the effectiveness of each part in our proposed model. We removed the siamese network from our proposed methods, resulting in just convolutional neural networks (CNNs). We implemented a single-head siamese network in which all sentences from all heads are put together into just one head. We also implemented two versions of a multi-head siamese network employing just one description or multiple descriptions, respectively.

The experimental results are shown in Table 5. The result shows that our proposed model gives the best performance when all of the modules are combined. Compared to CNN models, the performances are improved when the siamese network is added. Note that the siamese network contributes to accurate detection, since it captures the

Model	Detection Task							
	MDD		Bipolar		GAD		BPD	
MHS	F1	AUC	F1	AUC	F1	AUC	F1	AUC
w/depression	**89.6**	**93.8**	89.4	92.7	89.5	93.4	89.8	93.5
w/bipolar	88.2	92.4	**90.4**	**93.4**	90.4	93.2	88.8	91.8
w/anxiety	88.5	92.7	89.2	93.2	**91.5**	**94.3**	88.9	92.9
w/bpd	88.3	92.4	89.3	92.5	88.8	92.8	**90.8**	**94.0**

Table 6: The results of four mental illness detection tasks. The notation *w/(mental illness)* indicates the model takes symptom descriptions of the specific *mental illness* as input, respectively.

Model		RSDD		eRisk	
		F1	AUC	F1	AUC
Train: r/depression	BERT	35.7	50.8	52.3	78.1
	XLNet	34.9	50.5	52.8	78.5
	RoBERTa	37.4	51.6	52.5	78.3
	GPT-2	37.8	51.7	53.2	78.4
	PHQ9	37.2	51.5	53.3	78.8
	MHS	**38.6**	**52.0**	**54.9**	**79.5**

Table 7: The results of evaluation across the other dataset. Due to the uneven distribution of data, we report the weighted F1 scores for each test.

Model		Test: Target					
		Bipolar		GAD		BPD	
		F1	AUC	F1	AUC	F1	AUC
Train: MDD	Feature	54.0	69.1	49.5	66.6	55.2	69.8
	BERT	62.0	73.7	51.7	67.8	60.9	72.8
	XLNet	65.2	75.4	51.3	67.6	60.5	72.6
	RoBERTa	65.1	75.6	58.6	71.6	64.9	75.4
	GPT-2	65.2	75.7	**59.6**	72.1	62.6	73.5
	MHS w/depression	66.7	76.6	55.5	69.8	60.2	72.6
	MHS w/(=Target)	**76.6**	**85.4**	**59.6**	**72.2**	**67.5**	**77.3**

Table 8: The results of evaluation across the other mental disorders.

symptom-related features by comparing target texts with symptom descriptions. In addition, the performances are also improved when employing a multi-head rather than a single-head. It implies that individually training each symptom yields better results than training all symptoms together, as each symptom has unique features. Compared to learning from only one description per head, the performance of learning from multiple descriptions is improved. It may be due to each head learning further about the symptom through various sentences, covering distinct aspects of each symptom.

5.2 Contribution of Symptom Descriptions

To assess the effectiveness of symptom descriptions in detecting the presence of symptoms, we measure their performance by replacing the descriptions of symptoms with those of other mental disorders. The results are shown in Table 6. We carried out four mental disorder detection tasks using four models, each utilizing symptom descriptions of four distinct mental disorders as inputs.

The models exhibit optimal performance when the input symptom description corresponds to the target mental disorder. It suggests that, by providing the model with accurate and appropriate symptom descriptions, MHS can learn effectively to identify the specific features associated with a particular mental disorder. This also implies that MHS can identify and utilize the nuanced distinctions in the characteristics of each symptom, leading to enhanced performance in detection.

5.3 Cross-domain Test

In order to investigate the flexibility of MHS, we evaluated its performance across datasets and other mental disorders.

Dataset Transferability Given that the ability to generalize to new and unseen data platforms is a crucial aspect of mental illness detection models (Harrigian et al., 2020), we evaluate their performance across different datasets. We selected two datasets, RSDD (Yates et al., 2017) and eRisk2018 (Losada et al., 2019), to evaluate cross-dataset transfer. Unlike our Reddit dataset (Subsection 4.1), sourced from communities specific to certain mental illnesses, RSDD and eRisk2018 data are based on user self-reports, resulting in data that is different from and potentially unseen by the Reddit dataset. We trained each model using the Reddit train dataset and evaluated its performance on the test sets of RSDD and eRisk2018, respectively.

As shown in Table 7, MHS outperforms all strong baselines over all datasets. The improved performance of MHS compared to GPT-2, a strong few-shot learner, is likely due to its ability to leverage domain-specific knowledge. The higher generalizability of MHS compared to PHQ9 is likely attributed to its end-to-end architecture, which allows for direct learning of symptom features from data, as opposed to PHQ9's reliance on pre-defined symptom patterns.

Domain Transferability As suggested by some researchers (Aich and Parde, 2022; Uban et al., 2022), we evaluated the transferability of MHS across other mental disorders by training the models on a depression dataset and testing on other mental disorder datasets (see Table 8). The results of the experiments indicate that MHS significantly outperforms all relevant baselines, particularly when it utilizes symptoms that match the target mental disorder. This suggests that the transferability of

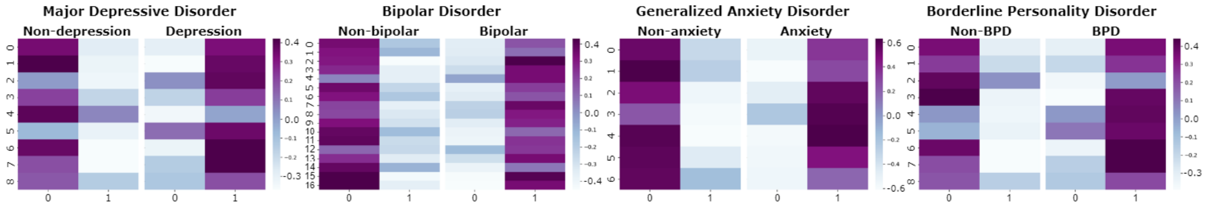

Figure 2: Examples of weights learned during the training process for each task. Each row represents a distance computed by each head, indicating the particular knowledge of the related symptoms.

the model can be significantly enhanced by simply replacing symptom descriptions. This also implies that it may be feasible to develop a model that can classify texts related to various other mental disorders if the symptoms of those disorders are provided appropriately.

5.4 Interpretation

Using our model, we can interpret the detected results by analyzing their representations of learned weights and distance values. In order to see if our model properly learned symptom-related knowledge from a few descriptions and identified similar stories from the target texts, we looked into the learned weights produced by the last step of our model, the fully connected layer. To show the effectiveness of MHS, we visualize the examples of learned weights from training steps in Figure 2. The color scale represents the strength of the learned weights (i.e., the distance values of each head). Each row represents heads, indicating each symptom referring to Table 1, and each column represents the labels. We observe a clearly contrasting pattern in the distance weights for each task.

We could also identify which symptoms are mainly activated or not by investigating the learned weights during the training process. For example, in detecting MDD-related texts, most of the symptoms have higher weights than depression. It suggests that most of the symptoms give rise to a major role during the detection process.

An important criterion in diagnosing a mental illness by experts is the number of expressed symptoms. The number of symptoms must exceed a certain number to be diagnosed as a corresponding mental illness. In order to see if the human-level diagnostic process works in our model as well, we looked into the number of salient symptoms in true-positive samples. We calculated percentiles from the similarity scores for each symptom in the true-positive samples from test sets, and set the threshold by 70% of the percentile. Then, when exceeding the threshold set by the criterion, the symptom was selected as a prominent feature in the text. We present the distribution of the numbers of salient symptoms and their averaged probabilities of the final output from test sets of detecting MDD-related texts in Figure 3.

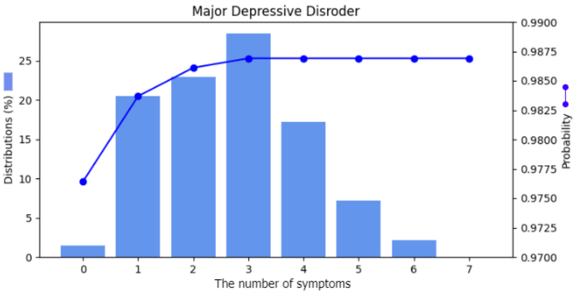

Figure 3: The number of salient symptoms and probability of the final output from true-positive samples in MDD detection.

In our model, the average probability is relatively low when there are fewer than three symptoms, but for three symptoms or more, our model makes a decision with high confidence at a similar level. It suggests that MHS also detects mental disorder-related texts with high confidence when the number of symptoms exceeds a specific number, the same as when humans diagnose. The criterion number being smaller in MHS may be due to the shorter length of social media texts, which may not fully convey the user's background and lifestyle.

5.5 Case Study

For the case study, we made an example based on the samples corresponding to each mental disorder in the psychology major textbook. We present example sentences for MDD and GAD (Table 9), and the model's predictions were correct in both cases. We set the same threshold as shown in Figure 3. The dominant symptoms predicted by the model are D0 (*depressed mood*), D1 (*diminished interest*), and D8 (*suicidal ideation*), for MDD, and A1 (*difficult to control the worry*), A2 (*irritability*), and A3 (*easily fatigued*), for GAD. In the case of D0

No.	Example	Expected Symptoms
1. (MDD)	Whenever I wake up in the morning, I hate myself, and I want to commit suicide. I didn't have any friends to hang out with because I did not need to make friends actively when I went to school. The only reason I am not committing suicide is I don't want my parents to cry.	**D0 (81%) D1 (80%)** D2 (25%) D3 (56%) D4 (10%) D5 (40%) D6 (47%) D7 (61%) **D8 (71%)**
2. (GAD)	I often feel anxious that something terrible is about to happen. For example, my husband will likely lose his job, or a family member will become ill or have an accident. I know these worries are unnecessary and excessive, but I can't stop worrying. I'm always nervous, so I feel exhausted even if I do nothing.	A0 (41%) **A1 (72%) A2 (75%) A3 (79%)** A4 (48%) A5 (34%) A6 (37%)

Table 9: Examples of texts related to MDD and GAD, respectively, and the corresponding symptoms that the models provide for interpretation. The notations of each symptom are referred to in Table 1.

and D1 in MDD, our model captures the feature related to the symptom, despite the absence of the term '*depress*' or '*interest*'. These cases support the assumption that our model can detect and interpret when symptoms of a particular mental illness are prominent in text.

6 Conclusion

In this paper, we proposed a simple but flexible model for detecting texts containing contents of mental health problems. Our model outperformed the state-of-the-art models and achieved human-interpretable results over symptoms regarding mental disorders. The proposed model demonstrates an exceptional ability to utilize domain knowledge as it is designed to capture relevant features from texts directly. Experimental results also indicate that MHS can quickly adapt to other mental disorder domains by simply replacing symptom descriptions. The scope of this paper was limited to the investigation of four mental disorder detection tasks. Nevertheless, this approach can be extended to other mental health conditions as long as the symptom-relevant questionnaires are provided accordingly.

Limitations

It should be noted that, as our model and the baseline models in this study were trained using texts from social media and the experiments were conducted on online text, the results may not accurately reflect the performance in a clinical setting. A proper diagnosis by clinical experts necessitates a comprehensive analysis of various factors, including the number of manifested symptoms, the onset and history of symptoms, developmental background, lifestyle, and recent life changes, in order to gain a comprehensive understanding of the patient's condition. However, it is still challenging to capture detailed information such as personal secrets through online text, as these texts are often composed of fragments of daily life, episodic experiences, and emotive expressions rather than providing a comprehensive view of an individual's life. Despite the domain-specific limitations imposed by the fragmentary text, we hope that our model may still serve as a valuable aid for clinical experts in their decision-making process. Furthermore, future research should aim to move beyond predicting psychological symptoms and disorders solely based on linguistic styles and expressions, and instead seek to uncover the underlying features that contribute to these expressions as our model does.

Ethics Statement

Since privacy concerns and the risk to the individuals should always be considered, especially using social media data, we have employed mechanisms to avoid any harmful and negative consequences of releasing our model. To this end, we removed individually identifiable information such as user names, user IDs, or e-mail addresses. We also removed any URLs from our data not to be trained on such personal information in our model. As for the use of open datasets in this work, we used them in accordance with guidelines that allow their use within the established usage policy. Especially we ensure that no attempts can be made to establish contact with specific individuals or deanonymize users in the datasets.

Our paper may contain direct references to specific disorders or diseases (such as psychiatric patients, Siamese, or names of mental disorders) and expressions that could be considered offensive to particular individuals. We want to emphasize that these expressions are used solely for the purpose of academic discourse and are not intended to be disrespectful or offend anyone.

In addition, our proposed model is not intended to label or stigmatize individuals online but rather to serve as a warning system for potential threats to personal well-being and public health. It is important to note that even if this model identifies potential mental illnesses and symptoms, it should not be considered a definitive diagnosis. Still, the model provides an indication of the likelihood of a disorder; it should be used as a reference for self-diagnose and in consultation with a mental health expert for an official diagnosis. An official diagnosis and results require consultation with medical and psychological experts, and this system aims at serving as an aid in the diagnostic process. We make our implementation code publicly available for research purposes, and we hope it will be used to improve the lives of individuals suffering from mental illnesses.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) (No. RS-2023-00208054, A multi-modal abusive language detection system and automatic feedback with correction) grant funded by the Korean government.

References

Ankit Aich and Natalie Parde. 2022. Are you really okay? a transfer learning-based approach for identification of underlying mental illnesses. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 89–104, Seattle, USA. Association for Computational Linguistics.

Kristen Allen, Shrey Bagroy, Alex Davis, and Tamar Krishnamurti. 2019. ConvSent at CLPsych 2019 task a: Using post-level sentiment features for suicide risk prediction on Reddit. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 182–187.

American Psychiatric Association. 2013. *Diagnostic and statistical manual of mental disorders (5th ed.)*. VA: American Psychiatric Association, Arlington.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758.

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

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*.

Munmun De Choudhury, Sanket S Sharma, Tomaz Logar, Wouter Eekhout, and René Clausen Nielsen. 2017. Gender and cross-cultural differences in social media disclosures of mental illness. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 353–369.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Anca Dinu and Andreea-Codrina Moldovan. 2021. Automatic Detection and Classification of Mental Illnesses from General Social Media Texts. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 358–366.

Johannes C Eichstaedt, Robert J Smith, Raina M Merchant, Lyle H Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A Asch, and H Andrew Schwartz. 2018. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208.

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*, pages 1–16, New York, NY, USA. Association for Computing Machinery.

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.

Molly Ireland and Micah Iserman. 2018. Within and between-person differences in language used across anxiety support and neutral reddit communities. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 182–193.

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.

Zhengping Jiang, Jonathan Zomick, Sarah Ita Levitan, Mark Serper, and Julia Hirschberg. 2021. Automatic detection and prediction of psychiatric hospitalizations from social media posts. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 116–121, Online. Association for Computational Linguistics.

Efsun Sarioglu Kayi, Mona Diab, Luca Pauselli, Michael Compton, and Glen Coppersmith. 2017. Predictive linguistic features of schizophrenia. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 241–250, Vancouver, Canada. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Prasadith Kirinde Gamaarachchige and Diana Inkpen. 2019. Multi-task, multi-channel, multi-input learning for mental illness detection using social media text. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, Hong Kong. Association for Computational Linguistics.

Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille.

Atharva Kulkarni, Amey Hengle, Pradnya Kulkarni, and Manisha Marathe. 2021. Cluster Analysis of Online Mental Health Discourse using Topic-Infused Deep Contextualized Representations. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, pages 83–93.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021. Gpt understands, too. *arXiv preprint arXiv:2103.10385*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

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.

Kate Loveys, Jonathan Torrez, Alex Fine, Glen Moriarty, and Glen Coppersmith. 2018. Cross-cultural differences in language markers of depression online. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 78–87, New Orleans, LA. Association for Computational Linguistics.

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, Minneapolis, Minnesota. 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.

Ankit Murarka, Balaji Radhakrishnan, and Sushma Ravichandran. 2021. Classification of mental illnesses on social media using RoBERTa. In *Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis*, pages 59–68, online. Association for Computational Linguistics.

Thong Nguyen, Andrew Yates, Ayah Zirikly, Bart Desmet, and Arman Cohan. 2022. Improving the generalizability of depression detection by leveraging clinical questionnaires. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8446–8459, Dublin, Ireland. Association for Computational Linguistics.

Sachin R Pendse, Kate Niederhoffer, and Amit Sharma. 2019. Cross-Cultural Differences in the Use of Online Mental Health Support Forums. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–29.

Daniel Preoţiuc-Pietro, Maarten Sap, H Andrew Schwartz, and Lyle Ungar. 2015. Mental Illness Detection at the World Well-Being Project for the CLPsych 2015 Shared Task. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology (CLPsych)*, pages 40–45.

Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying LMs with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5203–5212, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Hoyun Song, Jinseon You, Jin-Woo Chung, and Jong C. Park. 2018. Feature attention network: Interpretable depression detection from social media. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

Michael M Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2019. Detection of depression-related posts in reddit social media forum. *IEEE Access*, 7:44883–44893.

Ana Sabina Uban, Berta Chulvi, and Paolo Rosso. 2021. Understanding Patterns of Anorexia Manifestations in Social Media Data with Deep Learning. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 224–236.

Ana Sabina Uban, Berta Chulvi, and Paolo Rosso. 2022. Multi-aspect transfer learning for detecting low resource mental disorders on social media. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3202–3219, Marseille, France. European Language Resources Association.

Ning Wang, Fan Luo, Yuvraj Shivtare, Varsha D Badal, KP Subbalakshmi, Rajarathnam Chandramouli, and Ellen Lee. 2021. Learning Models for Suicide Prediction from Social Media Posts. *arXiv preprint arXiv:2105.03315*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Xingwei Yang, Rhonda McEwen, Liza Robee Ong, and Morteza Zihayat. 2020. A big data analytics framework for detecting user-level depression from social networks. *International Journal of Information Management*, 54:102141.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNET: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

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.

Zhiling Zhang, Siyuan Chen, Mengyue Wu, and Kenny Zhu. 2022. Symptom identification for interpretable detection of multiple mental disorders on social media. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9970–9985, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jinting Zhu, Julian Jang-Jaccard, Amardeep Singh, Paul A Watters, and Seyit Camtepe. 2021. Task-aware meta learning-based siamese neural network for classifying obfuscated malware. *arXiv preprint arXiv:2110.13409*.

A Experimental Setups

We implemented two feature-based models, support vector machine (SVM) and random forest (RF). We fine-tuned SVM with Gaussian kernel and set C to 100, and RF set max depth to 100. We employed BERT's vocabulary to train **BoW** models. For **Feature** models, we used a pre-trained sentiment classification model, and a Part-of-Speech Tagging model from the Huggingface library (Wolf et al., 2019). We fine-tuned the transformer baseline models employing the default settings from the Huggingface library: **BERT** (*bert-base-cased*), **XLNet** (*xlnet-base-cased*), **RoBERTa** (*roberta-base*), **GPT-2** (*gpt2*). For all experiments, we set the batch size as 8 and fine-tuned all models on a single 24GB GeForce RTX 3090 GPU. For the implementation of the **PHQ9** model, we follow the structure of the questionnaire-depression pair models by using the publicly available code from PHQ9[4] (Nguyen et al., 2022). We utilized the symptom patterns which are provided by Nguyen et al. (2022). We trained each of the models using all six randomly selected seeds, and all the models were trained for 3 epochs. We optimize the model parameters of all models with the Adam optimizer (Kingma and Ba, 2014). The learning rates for BERT, XLNet, and RoBERTa models were manually fine-tuned, choosing one out of {1e-05, 2e-05, 1e-06, 2e-06} that shows the best F1 score. The learning rate for GPT-2 was selected from {1e-05, 2e-05}, and for PHQ9, the learning rate was set to 1e-03, which was provided as an optimized hyperparameter.

B Comparison with Large Language Model

Recent developments in large language models (LLMs), such as GPT-3 (Brown et al., 2020), have demonstrated strong zero-shot performance across various NLP tasks. LLMs have the ability to achieve high performance without fine-tuning for downstream tasks, even with only zero or few examples, due to their large number of pre-trained parameters.

We experimented with obtaining results for the examples referred to in Table 9 by using GPT-3, a widely recognized LLM. To this end, we utilized instructional prompts by listing symptom descriptions for a specific mental illness. The examples of prompt input and the result are shown in Table 10. The experimental results show that the model successfully outputs the classification results in a sentence when given instructional prompts for a specific mental illness. However, the process of selecting symptoms appears to focus on identifying multiple symptoms rather than pinpointing a specific symptom with precision.

These examples are presented for demonstration purposes only, and the results may vary depending on the utilization of different prompt optimizations (Liu et al., 2021; Qin and Eisner, 2021). This aspect of research is beyond the scope of our current study; thus, there is room for further research to be conducted in future work.

C Details of Symptom Descriptions

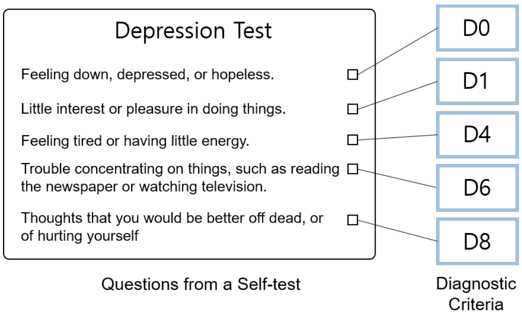
Depression Test
Feeling down, depressed, or hopeless. □
Little interest or pleasure in doing things. □
Feeling tired or having little energy. □
Trouble concentrating on things, such as reading the newspaper or watching television. □
Thoughts that you would be better off dead, or of hurting yourself □
D0
D1
D4
D6
D8
Questions from a Self-test
Diagnostic Criteria

Figure 4: An example mapping of questions into corresponding diagnostic criteria.

In this section, we present the symptom descriptions that were utilized in our current study. Table 11 shows the complete list of symptom descriptions. We used *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)* (American Psychiatric Association, 2013) as a reference for the symptom descriptions, as it provides comprehensive guidelines for identifying symptoms of various mental disorders. We also incorporated publicly available clinical questionnaires from online sources. Subsequently, under the guidance of a psychology researcher, we conducted a mapping process of the questions in the self-test to the corresponding diagnostic criteria, as depicted in Figure 4.

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

No.	Example
1. (MDD)	**Prompt:** These are symptom descriptions of major depressive disorder: D0. Depressed mood most of the day D1. Diminished interest or pleasure D2. Sleep disorders (insomnia or hypersomnia) D3. Changes in weight or appetite when not dieting D4. Fatigue or loss of energy D5. Feeling worthlessness or guilty D6. Diminished ability to think or concentrate D7. A slowing down of thought and a reduction of physical movement D8. Recurrent thoughts of death and suicidal ideation Please tell me if the user below has major depressive disorder and which symptoms does the user have? (choose from the above)
	Input: Whenever I wake up in the morning, I hate myself, and I want to commit suicide. I didn't have any friends to hang out with because I did not need to make friends actively when I went to school. The only reason I am not committing suicide is I don't want my parents to cry.
	Output: Yes, this user appears to have major depressive disorder and they have symptoms D0 (depressed mood most of the day), D1 (diminished interest or pleasure), D5 (feeling worthlessness or guilty), D6 (diminished ability to think or concentrate), and D8 (recurrent thoughts of death and suicidal ideation).
2. (GAD)	**Prompt:** These are symptom descriptions of generalized anxiety disorder: A0. Excessive anxiety and worry more than 6 monts A1. Difficult to control the worry The anxiety and worry are associated with followings: A2. Irritability A3. Being easily fatigued A4. Sleep disturbance A5. Difficulty concentrating or mind going black A6. Muscle tension Please tell me if the user below has generalized anxiety disorder and which symptoms does the user have? (choose from the above)
	Input: I often feel anxious that something terrible is about to happen. For example, my husband will likely lose his job, or a family member will become ill or have an accident. I know these worries are unnecessary and excessive, but I can't stop worrying. I'm always nervous, so I feel exhausted even if I do nothing.
	Output: Yes, the user has generalized anxiety disorder and they have the following symptoms: A0. Excessive anxiety and worry more than 6 months, A1. Difficult to control the worry, A2. Irritability, A3. Being easily fatigued, A4. Sleep disturbance, A5. Difficulty concentrating or mind going blank, A6. Muscle tension.

Table 10: Results for input example texts related to MDD and GAD, using GPT-3, respectively.

Major Depresive Disorder

D0: **Depressed mood most of the day, nearly every day.**
Feeling down, depressed, or hopeless.

D1: **Markedly diminished interest or pleasure in all, or almost all, activities most of the day, nearly every day.**
Little interest or pleasure in doing things.

D2: **Insomnia or hypersomnia nearly every day.**
Trouble falling or staying asleep, or sleeping too much.

D3: **Significant weight loss when not dieting or weight gain, or decrease or increase in appetite nearly every day.**
Poor appetite or overeating.

D4: **Fatigue or loss of energy nearly every day.**
Feeling tired or having little energy.

D5: **Feeling worthlessness or excessive or inappropriate guilt nearly every day.**
Feeling bad about yourself - or that you are a failure or have let yourself or your family down.

D6: **Diminished ability to think or concentrate, or indecisiveness, nearly every day.**
Trouble concentrating on things, such as reading the newspaper or watching television.

D7: **A slowing down of thought and a reduction of physical movement.**
Moving or speaking so slowly that other people could have noticed.

D8: **Recurrent thoughts of death, recurrent suicidal ideation without a specific plan, or a suicide attempt or a specific plan for committing suicide.**
Thoughts that you would be better off dead, or of hurting yourself.

Bipolar Disorder

Major Depressive Episode: D0-D8: Same as major depressive disorder.

Manic Episode:

M0: **A distinct period of abnormally and persistently elevated, expansive, or irritable mood and abnormally and persistently increased goal-directed activity or energy, lasting at least 1 week and present most of the day, nearly every day.**
Do you ever experience a persistent elevated or irritable mood for more than a week?

M1: **Increase in goal-directed activity or psychomotor agitation (i.e., purposeless non-goal-directed activity).**
Do you ever experience persistently increased goal-directed activity for more than a week?

M2: **Inflated self-esteem or grandiosity.**
Do you ever experience inflated self-esteem or grandiose thoughts about yourself?

M3: **Decreased need for sleep (e.g., feels rested after only 3 hours of sleep).**
Do you ever feel little need for sleep, feeling rested after only a few hours?

M4: **More talkative than usual or pressure to keep talking.**
Do you ever find yourself more talkative than usual?

M5: **Flight of ideas or subjective experience that thoughts are racing.**
Do you experience racing thoughts or a flight of ideas?

M6: **Distractibility (i.e., attention too easily drawn to unimportant or irrelevant external stimuli), as reported or observed.**
Do you notice (or others comment) that you are easily distracted?

M7: **Excessive involvement in activities that have a high potential for painful consequences.**
Do you engage excessively in risky behaviors, sexually or financially?

Anxiety Disorder

A0: **Excessive anxiety and worry, occurring more days than not for at least 6 months, about a number of events or activities.**
Do you worry about lots of different things? Do you worry about things working out in the future?
Do you worry about things that have already happened in the past? Do you worry about how well you do things?

A1: **The individual finds it difficult to control the worry.**
Do you have trouble controlling your worries? Do you feel jumpy?

A2: **The anxiety and worry are associated with irritability.**
Do you get irritable and/or easily annoyed when anxious?

A3: **The anxiety and worry are associated with being easily fatigued.**
Does worry or anxiety make you feel fatigued or worn out?

A4: **The anxiety and worry are associated with sleep disturbance (difficulty falling or staying asleep, or restless, unsatisfying sleep).**
Does worry or anxiety interfere with falling or staying asleep?

A5: **The anxiety and worry are associated with difficulty concentrating or mind going blank.**
Does worry or anxiety make it hard to concentrate?

A6: **The anxiety and worry are associated with muscle tension.**
Do your muscles get tense when you are worried or anxious?

Borderline Personality Disorder

B0: **A pattern of unstable and intense interpersonal relationships characterized by alternating between extremes of idealization and devaluation.**
My relationships are very intense, unstable, and alternate between the extremes of over idealizing and undervaluing people who are important to me.

B1: **Recurrent suicidal behavior, gestures, or threats, or self-mutilating behavior.**
Now, or in the past, when upset, I have engaged in recurrent suicidal behaviors, gestures, threats, or self-injurious behavior such as cutting, burning, or hitting myself.

B2: **Identity disturbance: markedly and persistently unstable self-image or sense of self.**
I have a significant and persistently unstable image or sense of myself, or of who I am or what I truly believe in.

B3: **Affective instability due to a marked reactivity of mood.**
My emotions change very quickly, and I experience intense episodes of sadness, irritability, and anxiety or panic attacks.

B4: **Inappropriate, intense anger or difficulty controlling anger.**
My level of anger is often inappropriate, intense, and difficult to control.

B5: **Transient, stress-related paranoid ideation or severe dissociative symptoms.**
I have very suspicious ideas, and am even paranoid or I experience episodes under stress when I feel that I, other people, or the situation is somewhat unreal.

B6: **Impulsively in at least two areas that are potentially self-damaging (e.g., spending, sex, substance abuse, reckless driving, binge eating).**
I engage in two or more self-damaging acts such as excessive spending, unsafe and inappropriate sexual conduct, substance abuse, reckless driving, and binge eating.

B7: **Frantic efforts to avoid real or imagined abandonment.**
I engage in frantic efforts to avoid real or imagined abandonment by people who are close to me.

B8: **Chronic feelings of emptiness.**
I suffer from feelings of emptiness and boredom.

Table 11: The complete list of collected sentences for each head. The diagnostic criteria, sourced from DSM-5, are shown in bold, and questions from clinical questionnaires are underlined.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Yes, in the "Limitation" section

☑ A2. Did you discuss any potential risks of your work?
Yes, in the "Ethics statement" section

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Yes, the paper's main claims are provided in the 1. Introduction section.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Yes, in 3. Methodology.

☑ B1. Did you cite the creators of artifacts you used?
Yes, in 2. Related work

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No, the codes will be publicly available after the reviewing process.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Yes, we discuss about the possible problems in the "Ethics statement" section.

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Yes, it is also discussed in "Ethics statement"

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Yes, in section 4.1 datasets

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Yes, in section 4.1 datasets

C ☑ Did you run computational experiments?

Yes, in Section 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Yes, in section 4, and Appendix

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Yes, in section 4, and Appendix

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Yes, in section 4 and 5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Yes, in section 4

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.