Learning from Mental Disorder Self-tests: Multi-head Siamese Network for Few-shot Knowledge Learning

Anonymous NAACL submission

Abstract

Social media is one of the most highly sought resources to analyze characteristics of the language by its users. In particular, many researchers utilized various linguistic features to identify users with mental disorders. However, generalizing linguistic features of such psychiatric patients is challenging since these features are apparently dependent on cultural or personal language habits. To address this challenge, we make use of the symptoms, which are shared properties of people with mental illness, concerning clinical contents rather than the ways of expressing them. In this paper, we aim to let our classification model identify informative features by training on knowledge about the symptoms. To this end, we propose a multi-head siamese network, which captures informative features based on the knowledge of mental illness symptoms and compares them to those of target text to be classified. The model is designed to learn the required knowledge by reading just a few questions from self-tests, and to identify similar stories from social media texts. Experimental results demonstrate that our model achieves improved performance as well as human-interpretable results for mental illness symptoms. A case study shows that our proposed model offers the possibility of automatic mental illness diagnosis, grounded on rational reasons.

1 Introduction

Mental disorders are usually accompanied by distinct symptoms, such as loss of interest or appetite, depressed moods, or excessive anxiety, which all hamper an individual's daily function. As these functional disruptions can often be manifested in social media, mental illness detection in social media is a field that has been studied extensively (Jiang et al., 2021; Kulkarni et al., 2021; Macavaney et al., 2021; Harrigian et al., 2020; Murarka et al., 2020; Gamaarachchige and Inkpen, 2019; Matero et al., 2019). Most researches proposed important features for mental illness detection, such as lexical features (Dinu and Moldovan, 2021; Jiang et al., 2020), sentiment or emotional aspects (Wang et al., 2021; Allen et al., 2019), or topic changes (Kulkarni et al., 2021; Tadesse et al., 2019).

These studies have been mainly grounded on the differences in linguistic features. However, it is challenging to generalize characteristics of psychiatric patients by such linguistic features since they are apparently dependent on subtle personal habits. For example, the manner people express their mental illness may vary to their resident culture (Loveys et al., 2018). To address this challenge, we focus on the shared and generalized properties of people with mental disorders. For this purpose, we propose to look into clinical contents rather than the way of expressing them, in detecting symptoms from texts. This is because, even though the language habits can differ individual by individual, patients share certain common symptoms. American Psychiatric Association (2013) compiled general and universal knowledge about such symptoms of mental disorders in DSM-5. We propose to make use of the knowledge about the symptoms to let our classification model learn informative features.

Several researchers have attempted to transfer such knowledge into their models for enhanced performance, exploiting graphical structures (Du et al., 2021; Hu et al., 2021; Cai and Lam, 2020), hierarchical structures (Zhang et al., 2021), or additional pre-training phases (Zhu et al., 2021b; Gururangan et al., 2020). In this paper, we use a more straightforward and intuitive approach, employing the siamese network, which adopts one-shot learning for domain-specific features (Koch et al., 2015). Using the siamese network, we also directly compare the input and the symptoms to find discriminative clues from texts. This process is motivated by that of humans who can quickly grasp a new idea, often by reading just a single explanation.

For example, when people are reading a depression self-test, they understand the questions, learn which symptoms are related to depression, and look back on their own behaviors, so as to self-diagnose their levels of depression. Inspired by this process, we propose a multi-head siamese network to let our model learn domain knowledge about symptoms of mental disorders from just a few sentences and identify manifested information from online posts. Additionally, by analyzing learned weights and distance values of each symptom, our model gives rise to human-understandable interpretations. We utilize the diagnostic criteria sourced from DSM-5 (American Psychiatric Association, 2013), and the self-tests that rephrase the symptoms colloquially. The self-tests are designed to be similar to social media texts by using day-to-day terms.

We evaluate the performance of our model on four mental disorder detection tasks, with data collected from online communities. We validate the performance of our model with respect to mental disorder detection and interpretability similar to real diagnosis. We show that our model shows performance as competitive as the state-of-the-art models, and yet learns appropriate knowledge with just a few examples. We also assess the effectiveness of multi-head siamese network in terms of its interpretability, which helps researchers to locate novel but important evidence. The implementation code and symptom-sentences will all be made publicly available[1].

2 Related Work

Social media are commonly used for mental health researches because of the ease of access for studies of various aspects of human behavior. Some researchers proposed such characteristics as differences in word usage between users with and without mental disorders (Dinu and Moldovan, 2021; Jiang et al., 2020; Tadesse et al., 2019), or in syntactic features (Yang et al., 2020; Ireland and Iserman, 2018; Kayi et al., 2018). Some studies address the differences between sentiments or emotional aspects (Wang et al., 2021; Allen et al., 2019; Gamaarachchige and Inkpen, 2019), or differences in topics (Kulkarni et al., 2021; Tadesse et al., 2019). Some researches also presented interpretable mental disorder detection methods based on linguistic features (Uban et al., 2021; Song et al., 2018). However, the linguistic characteristics may also vary to cultural or personal language habits (Loveys et al., 2018). Some studies employed strong Transformer based classifiers (Dinu and Moldovan, 2021; Jiang et al., 2020; Murarka et al., 2020), but they do not still deliver an expert-level analysis due to the lack of a wealth of knowledge about mental disorders.

Various efforts are made to transfer background knowledge or domain knowledge into their proposed models for enhanced performance. Some employed graphical structures to represent the semantic relations or additional knowledge (Du et al., 2021; Hu et al., 2021; Cai and Lam, 2020). Others made use of hierarchical structures, which require pre-defined hierarchical layers for knowledge representation learning (Zhang et al., 2021). Yet others attempted to transfer domain knowledge by an additional phase of pre-training with an in-domain corpus (Zhu et al., 2021b; Gururangan et al., 2020). However, all of these efforts require complicated steps in learning knowledge. In this paper, we use the siamese network (Koch et al., 2015), a straightforward and intuitive approach, exploited recently for simple networks (Chen and He, 2021; Zhu et al., 2021a). Its details are explained in the following section.

3 Multi-head Siamese Network

In order to simulate the process of mental disorder detection with domain knowledge, we designed our model based on the siamese network (Koch et al., 2015). As with the original siamese neural network, our model also contains symmetric twin networks with tied parameters. The symmetric twin networks are composed of multiple convolutional layers, and the outputs of each convolutional layer correspond to important features from input sentences. Employing the cosine similarity, we compute the distance values (d) between the two feature embeddings extracted from two inputs.

In addition, we apply multi-head few-shot learning to the original siamese network, repeating the distance calculation process by the number of related symptoms. Assuming that we have n symptoms for discriminating a mental disorder, we build a set of H heads for the mental disorder detection model as follows:

$$H = \{h_1, h_2, ..., h_n\} \quad (1)$$

Each head h_i represents domain knowledge regarding each symptom, which contains a number of questions from self-tests and an explanation of the

[1] https://xxx.yyy/zzz

corresponding symptom. For example, if h_1 has m sentences describing the symptom, we have a set of Q_{h_1} questions for a few-shot learning:

$$Q_{h_1} = \{q_1, q_2, ..., q_m\} \tag{2}$$

We describe the specifics of n symptoms for related mental disorders and the detailed structure of our model in the following subsections.

3.1 Symptom Descriptions

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.
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 symptoms.
	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.

In the present study, we focus on four mental disorders: major depressive disorder (MDD), bipolar disorder, anxiety disorder, 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 major depressive disorder, there are a total of 9 symptoms (D0-D8), so when constructing a depression detection model, there will be a total of 9 heads ($n(H_{dep.}) = 9$). As for bipolar, symptoms

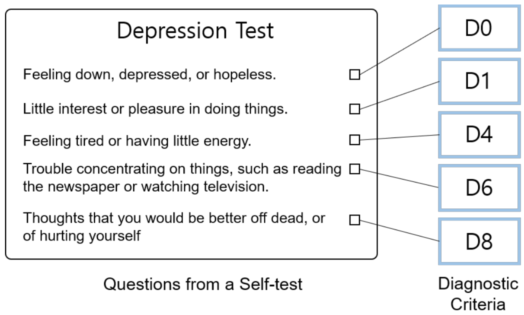

Figure 1: An example mapping of self-test questions into corresponding diagnostic criteria.

can be divided into depressive episodes (D0-D8) and manic episodes (M0-M7), with a total of 17 heads. The symptoms of bipolar disorder are the same as those of MDD.

Each head includes an explanation of diagnostic criteria and questions from self-tests corresponding to each symptom for few-shot learning. As a result, each head contains two or more sentences ($n(Q_h) \geq 2$). In the case of more than two related questions in the self-test, the corresponding head contains more than two sentences. Figure 1 shows the process of mapping the questions in the self-test to the corresponding diagnostic criteria. We collected the questions from the publicly available self-tests[2]. The process was conducted under the guidance of a psychology major researcher. The total / average number of sentences is 18/2 (MDD), 34/2 (bipolar), 18/2.6 (anxiety), and 18/2 (BPD). The complete list of collected sentences for each head is attached in Appendix A[3]. Each sentence from the heads will be another input to be compared to the input texts in the siamese network.

3.2 Model Structure

In this work, we aim to let our model learn knowledge about the mental illness symptoms, and identify salient features from input texts by comparing them with the learned knowledge. To this end, we propose a multi-head siamese network, as shown in Figure 2, which captures informative features based on the symptoms and compares them to a target text to be classified. With a given sequence of tokens as an input, our model tokenizes the input and obtains a sequence of embedding vectors (E_{input}) by employing pre-trained language model

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

[3]The supplementary materials (appendix) will be made publicly available with the code.

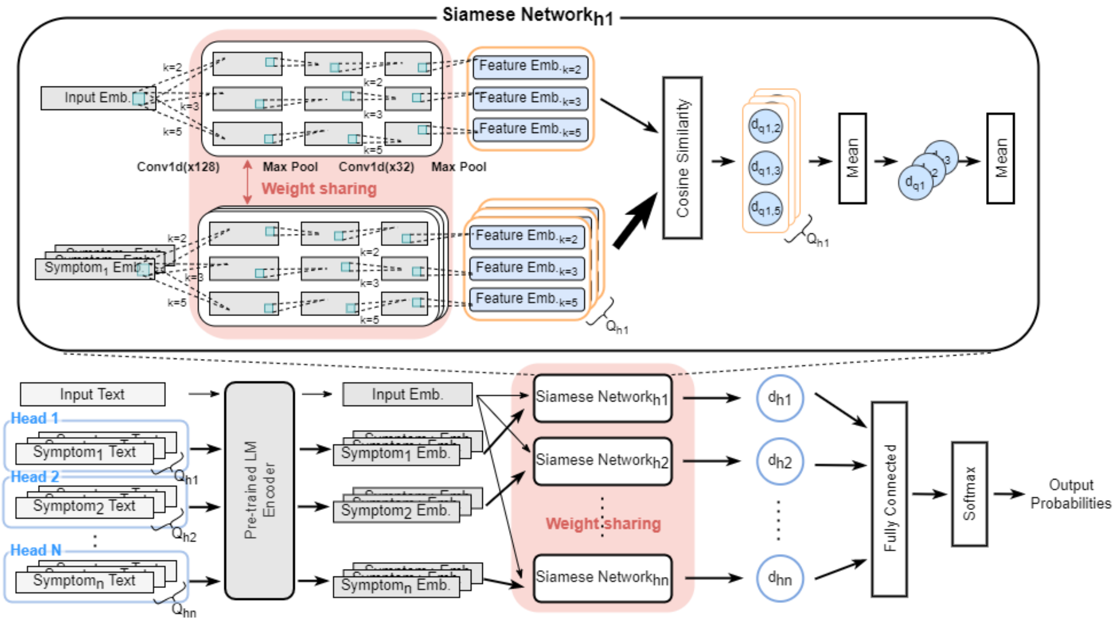

Figure 2: The model architecture of multi-head siamese network. d indicates the distance value computed by cosine similarity, and h_1 through h_n indicate the number of heads. Q_{h1} indicates the number of questions of h_1 for few-shot learning.

tokenizers, such as BERT tokenizer or RoBERTa tokenizer. We also get symptom embeddings (E_q) by encoding all sentences (Q_h) 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. Each channel contains two convolutional layers and two max-pooling layers. The final convolutional layer is flattened into a single vector, which is a feature embedding vector. As a result, we obtain three feature embedding vectors (F_{input}) from the input text:

$$F_{input,k} = Conv1d(E_{input})_k, (k = 2, 3, 5) \quad (3)$$

Through the same process, we also obtain feature embedding vectors from symptom texts (F_{qn}) from the n^{th} head as follows:

$$F_{q,k} = Conv1d(E_q)_k, (q \in Q_{hn}) \quad (4)$$

We compute the distances, in the range of [-1,1], through cosine similarity, comparing the input feature vector (F_{input}) and every sentence vector (F_q) prepared for few-shot learning:

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

$$d_{q,k} = sim(F_{input,k}, F_{q,k}), (q \in Q_{hn}) \quad (6)$$

Then we average the three distances (k=2, 3, 5) to get a single distance value between input text and a single sentence of the head:

$$d_q = mean(d_{q,2}, d_{q,3}, d_{q,5}), (q \in Q_{hn}) \quad (7)$$

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

$$d_{hn} = \frac{\sum d_q}{n(Q_{hn})}, (q \in Q_{hn}) \quad (8)$$

We iterate this process over the number of heads ($n(H)$). After the siamese network step, all distance values (d_{hn}) are stacked into a 1xn 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:

$$f = \mathbb{R}^N \rightarrow \mathbb{R}^2, \quad n(H) = N \quad (9)$$

$$o = f(D) = W^T \cdot D + b, \; (W \in \mathbb{R}^{N \times 2}) \quad (10)$$

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 will be covered in Section 5.3.

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, the average numbers of sentences and tokens, and the vocabulary size.

4 Experiments

4.1 Datasets and Evaluation

In order to evaluate our model, we constructed four datasets to detect each mental disorder. We sampled posts from Reddit[4], 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 mental illness group, and the other is from random subreddits as a non-illness group. 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), Anxiety disorder 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: Precision (Pre.), Recall (Rec.), and F1 score (F1) for the mental illness group. We report Accuracy (Acc.) of classification results. Also, we employ Area Under the Curve (AUC) to evaluate how much each model is capable of distinguishing between classes. The performance measure is reported by five-fold cross-validation and averaged after five runs.

4.2 Baselines and Experimental Setup

As for the baselines, we implemented two dictionary-based classifiers, support vector machine (SVM) and random forest (RF), and four pre-trained language based transformer models. We fine-tuned SVM with Gaussian kernel and C is set to 100, and RF where max depth is set to 100. We employed BERT's vocabulary to train dictionary-based models. We fine-tuned strong transformer baseline models employing the default settings from the Huggingface library (Wolf et al., 2019):

a. BERT (Devlin et al., 2019) is one of the most well-known baseline models (Jiang et al., 2020; Matero et al., 2019). We fine-tuned the *bert-base-cased* model.

b. ALBERT (Lan et al., 2019) has fewer parameters than the traditional BERT by two parameter reduction techniques. We fine-tuned the *albert-base-v2* model.

c. XLNET (Yang et al., 2019) is another strong baseline with a pre-trained language model (Dinu and Moldovan, 2021). We fine-tuned the *xlnet-base-cased* model.

d. 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., 2020). We fine-tuned the *roberta-base* model.

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 16 and 256, respectively, and fine-tuned over five epochs. We truncated each post at 256 tokens for all models. For each model, we manually fine-tuned the learning rates, choosing one out of {1e-5, 2e-5, 1e-6, and 2e-6} that shows the best F1 score. We report the average results over five-fold cross-validation runs on our dataset for the same pre-trained checkpoint.

4.3 Experimental Results

The experimental results of four mental illness detections for all baseline models and our proposed models are shown in Table 3. We report the mean for all metrics and the standard deviation (std.) of F1 scores on five-fold cross-validation tests. Our proposed model, the multi-head siamese network, is shown to outperform all the other strong baselines in all four tasks. On average, F1 is increased by 2.5% compared to the BERT and 0.9% compared to RoBERTa. AUC is increased by 2% compared to BERT and 1.1% compared to RoBERTa.

Table 4 shows the number of parameters for each model. Compared to the baseline models, the ad-

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

(a) Major depressive disorder detection

Model	Acc.	Pre.	Rec.	F1 (std.)	AUC
RF	89.9	**88.9**	63.0	73.7 (±0.3)	80.4
SVM	91.2	88.4	69.9	78.0 (±0.9)	83.6
BERT	94.2	85.8	89.0	87.3 (±0.2)	92.4
ALBERT	93.6	83.5	90.0	86.4 (±0.6)	91.3
XLNET	94.5	88.3	87.3	87.8 (±0.3)	92.4
RoBERTa	94.8	88.0	88.8	88.4 (±0.2)	92.7
ours†	94.8	85.3	**92.9**	88.9 (±0.4)	93.5
ours‡	**95.2**	86.9	92.4	**89.6** (±0.3)	**94.2**

(b) Bipolar disorder detection

Model	Acc.	Pre.	Rec.	F1 (std.)	AUC
RF	90.9	**94.5**	63.2	75.8 (±0.3)	81.1
SVM	90.2	77.3	79.0	78.2 (±0.8)	86.2
BERT	94.9	94.2	82.2	87.7 (±0.5)	90.3
ALBERT	94.5	90.4	84.5	87.3 (±0.4)	90.9
XLNET	94.9	86.2	**91.7**	88.9 (±0.4)	92.3
RoBERTa	95.5	92.9	86.1	89.4 (±0.3)	92.1
ours†	95.3	91.2	87.5	89.2 (±0.3)	92.3
ours‡	**95.8**	92.4	88.6	**90.4** (±0.3)	**93.3**

(c) Anxiety disorder detection

Model	Acc.	Pre.	Rec.	F1 (std.)	AUC
RF	91.7	**93.1**	64.6	76.3 (±0.4)	81.7
SVM	92.9	86.4	80.9	83.3 (±1.2)	88.5
BERT	95.3	91.2	86.0	88.5 (±0.5)	91.9
ALBERT	95.1	90.9	84.6	87.6 (±0.6)	91.2
XLNET	95.7	91.4	88.4	89.8 (±0.4)	93.2
RoBERTa	95.8	90.0	**91.7**	90.3 (±0.4)	93.4
ours†	95.8	89.9	90.8	90.3 (±0.4)	93.9
ours‡	**96.2**	92.0	91.0	**91.5** (±0.5)	**94.3**

(d) Borderline personality disorder detection

Model	Acc.	Pre.	Rec.	F1 (std.)	AUC
RF	90.3	90.8	61.3	73.2 (±0.3)	79.8
SVM	93.4	89.8	78.2	83.6 (±0.6)	88.9
BERT	95.0	85.7	**92.4**	88.9 (±0.3)	93.2
ALBERT	94.9	86.1	91.3	88.6 (±0.3)	93.2
XLNET	95.6	**92.9**	86.1	89.4 (±0.2)	92.3
RoBERTa	95.7	88.9	91.8	90.3 (±0.2)	93.3
ours†	95.7	89.9	90.7	90.4 (±0.4)	**94.0**
ours‡	**95.9**	91.1	90.4	**90.8** (±0.3)	**94.0**

Table 3: Mental illness detection results on (a) major depressive disorder detection, (b) bipolar disorder detection, (c) anxiety disorder detection, and (d) borderline personality disorder detection. † indicates that the model uses the BERT embeddings, and ‡ means that the model uses RoBERTa embeddings. The best results are shown in bold, and the second-best results are underlined.

Model	#parameters
BERT	108,311,810
RoBERTa	124,647,170
ours w/bert	108,967,319
ours w/roberta	125,302,679

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-shot	94.9	87.3	90.2	88.7	93.2
+multi-head +few-shot	95.4	89.1	90.5	89.7	93.9
CNNs w/roberta emb.	94.6	89.5	85.3	87.3	91.2
+multi-head +few-shot	**95.7**	**90.3**	**90.8**	**90.5**	**94.0**

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

ditional 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 ours† (w/bert) is slightly better or shows little difference. It suggests that our proposed model, learning domain knowledge, achieves efficient performance improvement by adding just a small number of parameters.

Additionally, even the dictionary-based model shows quite good performance, achieving high precision but low recall, indicating that the dataset shows distinct characteristics of each subreddit. However, compared to the dictionary-based model, the performance of the models with pre-trained language is improved by a significant difference. It means that some samples cannot be classified by a specific keyword, and the performance can be improved depending on how well the samples are classified. Since the dictionary-based models are mainly based on linguistic features, it may be difficult to find clues of mental illnesses, depending on the variance of linguistic habits. On the other hand, our model performed better than the baselines because it is designed to capture salient features based on learned symptoms, covering a broad range of clinical contents. The detailed analysis of the performance improvement is shown in Secction 5.

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 which result 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, and we also implemented a one-shot multi-head siamese network just using diagnostic criteria for each head. We compared both BERT embedding models and RoBERTa embedding models.

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. In addition, the performances are also improved when employing a multi-head rather than

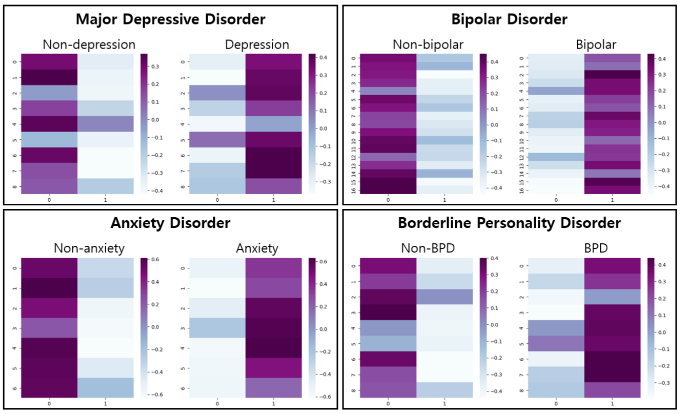

Figure 3: Examples of weights learned during the training process for each task. Each column represents a distance computed by each head, indicating the knowledge of the related symptoms.

		Trained Domain			
		depression	bipolar	anxiety	bpd
Target Domain	depression	**89.6**	88.9	88.1	87.8
	bipolar	89.6	**90.4**	88.7	88.3
	anxiety	89.5	90.4	**91.5**	88.8
	bpd	89.9	89.4	89.6	**90.8**

Table 6: The results of cross-domain tests. We report the F1 scores of each test.

a single-head. It is quite similar to a situation when experts diagnose mental illnesses, observing the number of symptoms from psychiatric patients. It suggests that training symptoms as separated knowledge is much more effective than learning all at once since each symptom is an independent factor. Compared to the one-shot method that learns only one sentence per head, the performance of few-shot is improved. It may be due to each head learning further about the symptom through various sentences, covering various aspects of each symptom. The performance is improved slightly when using RoBERTa embedding than when using BERT embedding. It suggests that plentiful embedding information may affect the performance.

5.2 Cross-domain Test

In order to see the exact reason for the performance improvement, we conducted a cross-domain test. The main goal of the cross-domain test is to see if the performance improvement was due to the learned contents, or whether the model itself compares several sentences. We also examine how cases with shared or similar symptoms between mental illnesses affect the performance.

We employed symptoms from the trained domain and used the input texts from the target domain. The results are shown in Table 6. The best performance, detecting each of the four target domains, shows up when training the same mental disorder knowledge. Bipolar disorder contains the most significant number of sentences about symptoms (in total, 34). However, when bipolar is employed as a trained domain, it could not show reasonable performance on the other domains. This suggests that training on the appropriate knowledge is required for enhanced performance with our model.

MDD and bipolar disorder share some symptoms, or the major depressive episodes. The result also shows good performance even after learning across the different domains. This implies that it may be possible to implement a model to classify various mental disorders into one model, if the symptoms of various mental illnesses are effectively assembled. We leave further details to future work.

5.3 Interpretation

Using our model, we can interpret the detected results by analyzing its representations of learned weights and distance values. In order to see if our model properly learns domain knowledge from a few sentences and identifies similar stories from the input texts, we looked into the learned weights produced by the last fully connected layer. To show our models' effectiveness, we visualize the examples of learned weights from training steps in Figure 3. The color scale represents the strength of the learned weights (i.e., 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 clear contrasting pattern in the distance weights for each task.

We could also identify which symptoms are

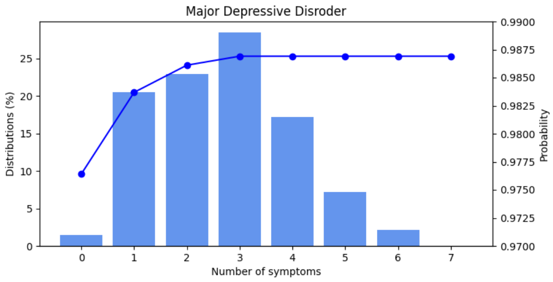

Figure 4: The number of salient symptoms and probability of the final output from true-positive samples in MDD detection.

mainly activated or not by investigating the learned weights during the training process. For example, in the MDD detection, most of the weights of symptoms give higher weights to the depression, except D4 (loss of energy). It suggests that most of the symptoms give rise to a major role during the detection. In the case of D4, we may improve the performance further by fine-tuning the symptom-related sentences.

An important criterion in diagnosing a mental illness by experts is the number of manifested symptoms. The number of symptoms must exceed a certain number to be diagnosed as a corresponding mental illness. For example, in the case of MDD, at least 5 out of 9 symptoms must be manifested to be diagnosed. 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 MDD detection in Figure 4.

In our model, the average probability is low when there are fewer than three symptoms, but when three symptoms or more, our model makes a decision with high confidence at a similar level. It suggests that our model also diagnoses a mental disorder when the number of symptoms exceeds a specific number, the same as when humans diagnose. The criterion number being smaller in our model may be due to the shorter length of social media texts.

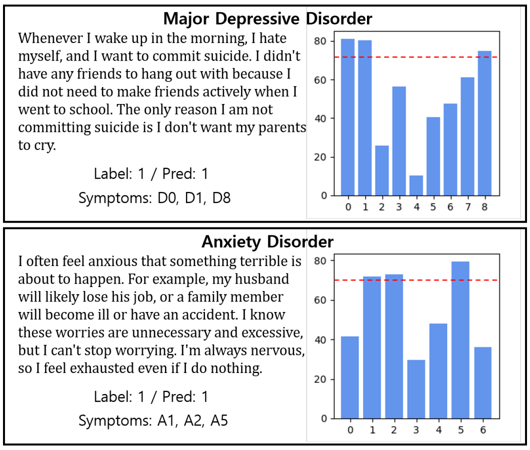

Figure 5: Examples with symptoms of corresponding mental disorder. The *label* indicates a gold standard, and the *pred* indicates the prediction of our model.

5.4 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 anxiety disorder (Figure 5), and the model's predictions were correct in both cases. We set the same threshold as shown in Figure 4. As for MDD, the salient symptoms predicted by the model are D0, D1, and D8, and for anxiety disorder, the prominent symptoms are A1, A2, and A5, and the model can identify most of the related terms in the text. In the case of D0 (depressed mood) and D1 (diminished interest or pleasure) in MDD, however, 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 disorder are prominent in text.

6 Conclusion

In this paper, we proposed a multi-head siamese network for mental disorder detection. Our model achieved improved performance as well as human-interpretable results over symptoms regarding mental disorders. We anticipate that the proposed model will provide an automatic mental illness diagnosis at the same level as human experts practice. In this study, we used social media texts. If we use medical data such as psychotherapy records, our model may turn out to be more prosperous in training symptoms. For cases such as bipolar or multi-disorder detection, it would be worth considering a hierarchical structure in the multi-head siamese network. We leave it for future work.

References

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: Dsm-5. *Arlington, VA: Author.*

Deng Cai and Wai Lam. 2020. Graph transformer for graph-to-sequence learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7464–7471.

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.

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*.

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.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2021. ExCAR: Event Graph Knowledge Enhanced Explainable Causal Reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2354–2363.

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)*, pages 54–64.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2020. On the state of social media data for mental health research. *arXiv preprint arXiv:2011.05233*.

Linmei Hu, Tianchi Yang, Luhao Zhang, Wanjun Zhong, Duyu Tang, Chuan Shi, Nan Duan, and Ming Zhou. 2021. Compare to The Knowledge: Graph Neural Fake News Detection with External Knowledge. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 754–763.

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.

Zheng Ping 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.

Efsun Sarioglu Kayi, Mona Diab, Luca Pauselli, Michael Compton, and Glen Coppersmith. 2018. Predictive linguistic features of schizophrenia. *arXiv preprint arXiv:1810.09377*.

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.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

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*.

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.

Sean Macavaney, Anjali Mittu, Glen Coppersmith, Jeff Leintz, and Philip Resnik. 2021. Community-level research on suicidality prediction in a secure environment: Overview of the CLPsych 2021 shared task. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 70–80.

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.

Ankit Murarka, Balaji Radhakrishnan, and Sushma Ravichandran. 2020. Detection and Classification of mental illnesses on social media using RoBERTa. *arXiv preprint arXiv:2011.11226*.

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*.

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.

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.

Fuxiang Zhang, Xin Wang, Zhao Li, and Jianxin Li. 2021. TransRHS: a representation learning method for knowledge graphs with relation hierarchical structure. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 2987–2993.

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

Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. 2021b. Topic-driven and knowledge-aware transformer for dialogue emotion detection. *arXiv preprint arXiv:2106.01071*.

A Appendix

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 7: The complete list of collected sentences for each head. The diagnostic criteria, sourced from DSM-5, are shown in bold, and questions from self-tests are underlined.