Detection of Multiple Mental Disorders from Social Media with Two-Stream Psychiatric Experts

Anonymous ACL submission

Abstract

Existing Mental Disease Detection (MDD) research largely studies the detection of a single disorder, overlooking the fact that many mental diseases might occur in tandem. Many approaches are not backed by domain knowledge (e.g., psychiatric symptoms) and thus fail to give interpretable results. We propose PsyEx, an MDD framework that is capable of learning the shared clues of all diseases, while also capturing the specificity of each single disease. The two-stream architecture which simultaneously processes text and symptom features can combine the strength of both modalities and offer knowledge-based explainability. Experiments on the detection of 7 diseases show that PsyEx can boost detection performance by more than 10%, especially in relatively rare classes.

1 Introduction

Mental Disease[1] Detection (MDD) is of great practical value and social benefits, because mental disorders can greatly affect sufferers' life quality (Dreisbach et al., 2019). Lots of practices (Coppersmith et al., 2015; Mowery et al., 2017) indicate that people's posts on social media, containing sufficient expressions about their feelings and symptoms, can be an informative data source for text-based automatic MDD, which aims to predict whether a person suffers from certain mental diseases.

However, traditional MDD methods (Yates et al., 2017; Trotzek et al., 2018) process every post in the user's posting history, which can include many irrelevant or distracting posts. To avoid these noises, some prior works try to extract key posts by clustering (Zogan et al., 2021) or semantic similarity (Zhang et al., 2022a), but these heuristics can still introduce erroneous posts, affecting the subsequent MDD results.

[1] In this work, we will use 'mental disorder' and 'mental disease' interchangeably.

Moreover, comorbidity of several mental disorders is common (Roca et al., 2009). For instance, 75% of depression patients in the surveyed population also suffer from anxiety disorder in their lifetime (Lamers et al., 2011). Some research (Adam, 2013) further suggests that mental diseases lie along a spectrum, hence it is quite usual for one person to develop symptoms of several related mental disorders at the same time, and be diagnosed with multiple diseases. However, detecting multiple mental diseases simultaneously in the scenario of comorbidity is still under-explored. Most existing works focus on the detection of a single common disorder, such as depression (Losada et al., 2017; Lee et al., 2021), ignoring the frequent comorbidity diagnosed in clinical practice.

To address these limitations, we aim to explore an approach to detect multiple mental disorders simultaneously in a comorbidity dataset, in which the diagnosed users can have one or more disorders. For simplicity, we refer to this detection task as **Multiple MDD** in this paper, and it can be viewed as a multi-label classification problem.

Intuitively, detecting multiple mental diseases can be challenging to resolve, as there are lots of overlapping clinical manifestations shared among different diseases, so the labels implicitly intersect with each other. Pioneering works on multiple MDD (Cohan et al., 2018; Sekulic and Strube, 2019) commonly yield unsatisfying results as they simply use a shared model architecture for all diseases, which may not be strong enough to distinguish between diseases. Zhang et al. (2022b) shows the effectiveness of implementing psychiatric symptom knowledge on multiple MDD, which is more interpretable and can outperform previous pure-text methods due to its better clinical grounding. However it detects each disease separately, ignoring the inner correlations between diseases, leading to unsatisfying results on rarer diseases like OCD.

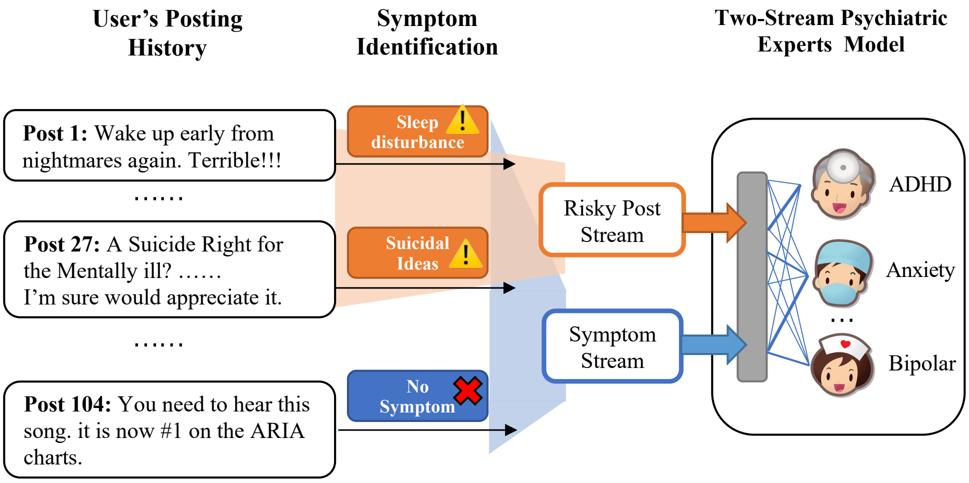

Figure 1: Psychiatric Experts Model with Symptom-based Risky Post Screening. Only posts highly related to psychiatric symptoms will be selected as diagnostic basis for further mental disease detection, which forms the risky post stream (the orange part). The symptom stream of a user contains all the symptom identification results from his/her whole posting history (the blue part), providing a more global view for the MDD model.

In this work, we propose **PsyEx** (**Psy**chiatric **Ex**perts), a multi-task learning framework that can simultaneously detect 7 mental disorders[2] with a shared backbone and disease-specific structures to leverage the common characteristics of all diseases while still being able to capture their distinctions. This framework consists of two phases (see Figure 1). We first utilize a symptom identification model to obtain *symptom stream*, as well as select posts that are highly relevant to symptoms as *risky post stream*[3]. These two streams are fed into the multiple MDD model which has 7 disease-specific psychiatric experts on top of a shared hierarchy network, so that the backbone can learn the shared knowledge between diseases, while each expert attending to one specific disease can focus on their own domain. Further, the symptom stream utilizes psychiatric knowledge explicitly, while this knowledge is implicitly embedded into the risky post stream as it is filtered by symptoms. Therefore, these two streams can complement each other and improve interpretability based on domain knowledge. Experiments show that our method can achieve SOTA multiple MDD performance across 7 diseases and bring significant improvement to rarer diseases on which the baselines struggled heavily. Our main contributions are:

- We successfully exploit psychiatric symptoms in this multiple MDD framework, including the symptom-based screening to facilitate a precise selection of risky posts, and the symptom stream to provide domain knowledge, together with a more holistic view of the entire posting history.
- We propose a two-stream Psychiatric Model for better multi-task learning of 7 mental disorders, which boosts the overall performance by better utilizing the distinctions and commonality among diseases.
- With the interpretability enabled by symptom and disease-specific experts, we analyze the decision-making process of PsyEx, and further study the contribution of each symptom to the detection of different diseases.

2 Approach

In this section, we introduce the proposed framework for multiple MDD, including symptom identification and a two-stream Psychiatric Experts Model. The symptom identification phase can generate the symptom stream and utilize a symptom-based risky post screening to select posts for the risky post stream.

2.1 Symptom-based Risky Posts Screening

Traditional MDD method processes every single post equally overlooking the fact that not every post

[2] The 7 mental disorders are: ADHD, Anxiety Disorder, Bipolar Disorder, Depression, Eating Disorder, OCD and PTSD. Brief introductions about these disorders are included in Appendix A.

[3] We call them "streams" here because both symptom features and text features are arranged in chronological order.

from a patient reveals useful information for detection. To facilitate the performance of the MDD model, as well as provide explainable diagnosis basis, we screen risky posts first, in particular with disease-dependent symptom information, and use the selected posts for multiple disease detection.

Inspired by clinical diagnosis procedures, we assume that posts reflecting *symptoms* of mental disorders suggest higher risks. Since healthy individuals rarely produce symptom-related content, posts with symptom indications could better separate patients from control users. Hence, unlike the prior heuristic methods discussed in Section 1, we implement a screening method based on the symptom features extracted by a supervised symptom identification model (Zhang et al., 2022b). This model can identify 38 symptom classes from 7 mental diseases with a Mental BERT-based encoder (Ji et al., 2022) and a linear classifier. It is trained on a large-scale, multi-disease annotated symptom identification dataset, based on the diagnostic criteria from DSM-5 (APA et al., 2013).

The symptom feature of each post is a 38-dimensional vector, where each dimension is the predicted probability of a certain symptom [4]. We can then estimate the *risky score* of a post with the *sum* of predicted probability among all symptoms. The top K posts with highest risky score among a user's whole posting history (containing N posts) will be selected as his/her risky posts. In practice, we set $K \ll N$, so the reduced input size enables high efficiency even with BERT-based language models in the following disease detection phase. Moreover, the symptom extracted by a supervised model is also more reliable to select risky posts than heuristic approaches, and we will show this in the experiments (Section 3).

2.2 Two-Stream Psychiatric Experts Model

Although MDD can be formulated as a typical text classification problem, traditional model structures are not appropriate for our multiple mental disease detection task, because they detect all the diseases using a single user representation, failing to model the distinctions of different diseases explicitly.

To alleviate this problem, we propose a Two-Stream Psychiatric Experts Model (See Figure 2), which can detect all the mental diseases simultaneously through multi-task learning while still capturing the nuances among them. It takes *risky post stream* and *symptom stream* obtained in symptom identification phase as input to combine the advantages of both modalities.

[4]These symptoms (e.g., *anxious mood, sleep disturbance, poor memory*) are carefully extracted from DSM-5, ensuring there is as little overlap as possible between them. We list all the symptoms in Table 10 in the Appendix.

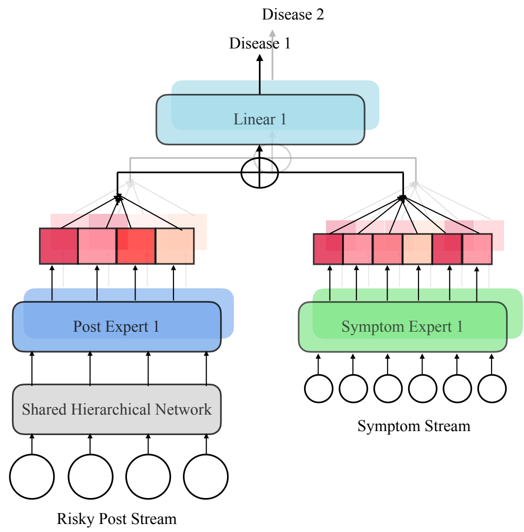

Figure 2: Illustration of the proposed model architecture, with 2 disease-specific experts on top of a shared network. Darkness of red square indicates attention strength, which is different for each disease.

Risky Post Stream Model The risky post stream "cherry-picks" K posts with highest symptom probability, providing more strong and concentrated signals of mental diseases. To better explore these signals and distinguish among multiple diseases, the model consists of two components: a single **shared hierarchical network** and D **task-specific attention layers** on top of the shared backbone, learning the special clues and features like "experts" in each disease.

In the shared part, we draw on the structure of Hierarchical Attention Network (HAN) (Yang et al., 2016) with post and user-level encoder to make better use of the posts' sequential structure. We employ a pre-trained BERT model as the post encoder. For words $\{w_1, w_2, ..., w_L\}$ in a post, the post representation p is,

$$p = BERT_{[CLS]}(w_1, w_2, ..., w_L) \qquad (1)$$

The user-level encoder utilizes a transformer (Vaswani et al., 2017) structure, modeling the relations between these posts $\{p_1, p_2, ..., p_K\}$, and produces updated post representations $\{p'_1, p'_2, ..., p'_K\}$.

In the disease-specific part, each disease has its own attention layer to get different attention distributions on the same post sequence. As such, these D attention layers can be considered as feature selectors from the shared network (Liu et al., 2019). Then we perform a weighted sum of the post representations according to attention score to get the distinctive user representations for each disease d.

$$\alpha_{k,d} = \frac{exp(W_d p'_k + b_d)}{\sum_{k'=1}^{K} exp(W_d p'_k + b_d)} \quad (2)$$

$$u_d = \sum_{k=1}^{K} \alpha_{k,d} p'_k \quad (3)$$

where W_d and b_d are both learnable parameters specifically for disease d.

Symptom Stream Model Symptom feature is critical for MDD, because it equips the detection model with explicit psychiatric knowledge. The symptom stream can be considered as an $N \times 38$ matrix, revealing the users' whole posting history in a lighter way, which is able to replenish the incomplete text stream and provide holistic information for the detection model. To better capture the unique features of different diseases, the symptom stream model is totally disease-specific, and its structure is the same as the corresponding part in risky post stream model. Similarly, each disease has its own "expert" to get different attention distributions on the same symptom sequence, which is intuitive because each mental disease has its own typical or unique symptoms for diagnosis (e.g., panic fear for Anxiety, intrusion for PTSD).

Finally, the disease-specific user representation from both stream are concatenated and fed into separate linear layers to get the binary predictions on whether the user suffers from certain mental disease.

The whole model is trained with the standard binary cross entropy loss, where the loss of all the tasks are averaged. We applied *loss masking* (Fonseca et al., 2020), so that for patients with at least one disease, we do not treat their absent disease labels as negative to alleviate the problem of potentially missing labels.

3 Experiments

In this section, we conduct experiments to (1) examine the performance of our two-stream PsyEx model compared with strong baselines; (2) validate the effectiveness of various design choices with ablation tests; (3) analyze the interpretability enabled by symptoms and disease-specific "experts".

3.1 Multiple MDD Dataset

We construct a multiple MDD dataset by reimplementing the data collection method of SMHD (Cohan et al., 2018). Users and posts were extracted from a publicly available Reddit corpus[5]. We select diagnosed users by detection patterns with a focus on high precision. The patterns consist of two components: one that matches a self-reported diagnosis (e.g., "diagnosed with"), and another that maps relevant keywords to the 7 mental diseases (e.g., "panic disorder" to Anxiety). A user is assigned a disease if one of its keywords occurs within 40 characters of the diagnosis pattern. Control users (i.e., healthy persons) are randomly sampled from those who never posted or commented in mental health related subreddits and never mentioned the name of 7 mental diseases (e.g., "bipolar", "PTSD") to avoid possible false positives. Similar to SMHD, we eliminate the diagnostic posts from the dataset to prevent the direct leakage of label, but retain those mental health related posts to allow the extraction of symptom-related features.

The final dataset consists of 5,624 diagnosed users and 17,209 control users. Each diagnosed user can have one or more disease labels, so we provide the distribution of the users' disease counts in Table 7 (Appendix B). The statistics show that 57% users in the dataset suffer from two or more kinds of mental disorders, and this comorbidity scenario is challenging for disease detection models. Moreover, due to the uneven distribution of different mental disorders in reality, the dataset is naturally unbalanced, with more users suffering from Depression and Anxiety, while fewer users with OCD and PTSD[6]. Consequently, the detection of rarer diseases can be even more difficult to resolve.

3.2 Methods of Comparison

For multiple MDD task, we mainly compared the proposed methods with 4 types of baselines: **TF-IDF+LR** (Cohan et al., 2018) is a representative traditional machine learning method which utilizes TF-IDF to extract textual features, followed by a Logistic Regression model for prediction. **BERT** is the reimplementation of the MDD model in Nguyen et al. (2022), which utilizes CNN of various kernel sizes on top of the sentence embeddings

[5] `https://files.pushshift.io/reddit/`

[6] See Table 6 in Appendix B for the exact number of users suffering from each disease.

Method	Depression	Anxiety	ADHD	Bipolar	OCD	PTSD	Eating	Avg. F1
TF-IDF+LR (Cohan et al., 2018)	76.75	75.45	68.97	76.56	40.0	44.02	30.0	58.82
BERT (Nguyen et al., 2022)	73.28	70.59	56.72	61.21	41.82	50.0	32.67	55.18
Symp (Zhang et al., 2022b)	81.06	81.99	70.3	81.75	65.12	64.41	61.54	72.31
HAN-GRU (Sekulic and Strube, 2019)	74.99	82.16	81.72	80.28	70.59	67.67	68.57	75.14
PsyEx	**87.89**	**89.84**	**84.40**	**91.58**	**81.69**	**81.69**	**85.71**	**86.12**

Table 1: Mental Disease Detection Results across 7 diseases, reporting F1 scores in binary setting. We order these diseases in descending order according to their number of patients in the dataset, making it easier to identify rarer classes.

Method	Depression	Anxiety	ADHD	Bipolar	OCD	PTSD	Eating	macro-F1	micro-F1	EM
TF-IDF+LR	52.16	32.52	29.9	43.14	13.79	9.76	0	25.9	38.09	77.37
BERT	56.99	47.03	31.86	16.59	0	0	0	21.78	39.38	75.32
Symp	62.99	**57.93**	49.57	51.85	18.18	0	0	34.36	53.05	76.65
HAN-GRU	59.4	44.88	53.81	62.2	0	0	0	31.47	50.47	77.55
PsyEx	**66.09**	54.97	**54.07**	**66.17**	**41.51**	**44.83**	**9.52**	**48.17**	**58.59**	**81.38**

Table 2: Mental Disease Detection Results across 7 diseases in mutli-label setting, reporting F1 score of each disease, micro-F1 and macro-F1 over all diseases, as well as exact match ratio (EM).

from pre-trained BERT as features to aggregate the information from users' posting list. **Symp** (Zhang et al., 2022b) uses the same CNN backbone, but further replace the BERT embedding features with symptoms features, and it establishes a strong baseline. **HAN-GRU** reimplements the hierarchical attention network for MDD proposed in Sekulic and Strube (2019), which utilizes Bidirectional GRU as encoders. More details like hyperparameter settings of the baseline experiments can be found in Appendix C.

3.3 Experimental settings

For PsyEx model, we select 16 high-risk posts during the screening process to form the risky post stream[7]. we utilize pre-trained bert-tiny[8] (in binary setting) or mental-bert[9] (in multi-label setting) as the basis of the post encoder. The user encoder is a 4-layer 8-head transformer encoder. We train with a batch size of 32 and set learning rate at $1e^{-5}$. We also employ early-stopping with a patience of 4 epochs according to validation performance to prevent overfitting.

Further, to infer the symptoms mentioned in the posts, we perform some pre-processing steps. First, we use *blingfire*[10] to split a post into sentences. Then, we use regular expressions to filter out the hyperlink format like "[anchor text] (web url)" and preserve the anchor text. Finally, we remove sentences like "[removed]".

3.4 Experiment Results

To be consistent with previous works (Cohan et al., 2018), we train and evaluate the models in both binary and multi-label setting.

3.4.1 Binary Setting Results

For the binary task, we only need to decide whether the user is suffering from a certain mental disease, so only users with this mental disorder plus all control users are selected to train and evaluate. The results in binary setting are shown in Table 1. We can see that our proposed *PsyEx* outperforms all the baseline methods including the strong *Symp* model, suggesting the advantage of our symptom-based risky post screening and two-stream psychiatric expert model. Further, owing to the multi-task learning and the shared knowledge between diseases, the detection effect of rarer classes (i.e., eating disorder, OCD, PTSD) is largely improved.

3.4.2 Multi-label Setting Results

In multi-label setting, we have to determine if and which mental diseases the user was diagnosed with, that is, the user can have one or more diseases at the same time, and all data is used for both training and evaluation. We show the multi-label results in Table 2, in which we evaluate these classifiers with a strict metric, exact match ratio, together with macro and micro F1 score which take partially correct into consideration.

This setting is challenging and underexplored in previous works mainly due to the complexity brought by comorbidity, as well as the various

[7] We further explore the impact of post number on the detection results in Figure 5 (Appendix D).

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

[9] https://huggingface.co/mental/mental-bert-base-uncased

[10] https://github.com/microsoft/BlingFire

overlapped manifestations among different mental disorders. Our PsyEx model shows significant superiority over other strong baselines, especially on the rarer classes, in which some classifiers can not even find a true positive sample.

3.4.3 Ablation Study

We examined the effectiveness of various design choices of the proposed Two-Stream PsyEx model with ablation tests in binary setting. Results are shown in Table 3.

Method	Avg. F1
Two-Stream PsyEx	**86.12**
w/o symp-stream	84.67
w/o multi-attn	83.68
w/o multi-task	85.40

Table 3: Ablation tests for the design choices of PsyEx, reporting F1 score averaged across 7 diseases. Results of each disease are shown in Table 8 (Appendix D).

First, we implement a **w/o symp-stream** model[11], which only preserves the risky post stream part to exhibit the effectiveness of symptoms. As shown in Table 3, the detection performance drops without symptom stream, indicating that symptoms can not only disentangle multiple diseases better with embedded domain knowledge, but also provide a global view of users' entire posting history.

Then, we examine the D disease-specific attention layers by implementing a model **w/o multi-attn**, in which all the diseases share a single attention head but are still trained simultaneously with multi-task learning and both streams are preserved as well. We can see that the performance is greatly harmed without multiple attention heads, illustrating the effectiveness of disease-specific "experts" to properly capture the characteristics of different mental diseases.

Moreover, the **w/o multi-task** method further trains a *w/o multi-attn* model for each disease separately without multi-task learning. We can notice a slight decrease on the detection performance even with much more model parameters, as we need to train 7 independent models for *w/o multi-task* method.

We also compare our symptom-based risky post screening method with other approaches in the literature. **Similarity** and **K-Means** both utilize a pretrained Sentence BERT (Reimers and Gurevych, 2019) to obtain sentence representations. The former one extracts key posts according to the cosine similarity between post and mental disease descriptions, and select K posts with the highest similarity score. The later one runs K-means clustering algorithm on the post embeddings and gets the K posts nearest to the cluster center as representative posts. **Last** simply selects the *last* K post as risky posts. Except for the different screening methods, we use the same model structure (i.e., our proposed PsyEx), and experiment results are listed in Table 4. It can be seen that symptom-based risky post screening outperforms all the heuristic approaches, especially *K-Means* and *Last*, highlighting the importance of a precise screening method, as there can be large amount of posts irrelevant to mental disorders in the users' posting history.

Screening Method	Avg. F1
Symptom-based (PsyEx)	**86.12**
Similarity (Zhang et al., 2022a)	82.47
K-Means (Zogan et al., 2021)	74.76
Last	56.74

Table 4: Ablation test for different risky post screening methods, reporting F1 score averaged across 7 diseases. Results of each disease are shown in Table 9 (Appendix D).

3.5 User-level Interpretability Analysis

We believe that the interpretability enabled by the usage of symptom and disease-specific "experts", can be manifested in two aspects:

- **Difference**: different symptoms can contribute differently to predicting different diseases.
- **Reasonableness**: the contribution proportion of symptoms to different diseases in the model is reasonable (i.e., is roughly consistent with the authoritative DSM-5 criteria (APA et al., 2013)).

Therefore, we provide a concrete user-level example to illustrate the decision-making process of PsyEx. The selected user suffers from three mental disorders, including *anxiety, bipolar* and *depression*. We apply the trained PsyEx model to his/her posting history, and obtain the attention score matrix of symptom stream, which is $D \times N$,

[11] "w/o" means without, so symptom stream is removed in this model.

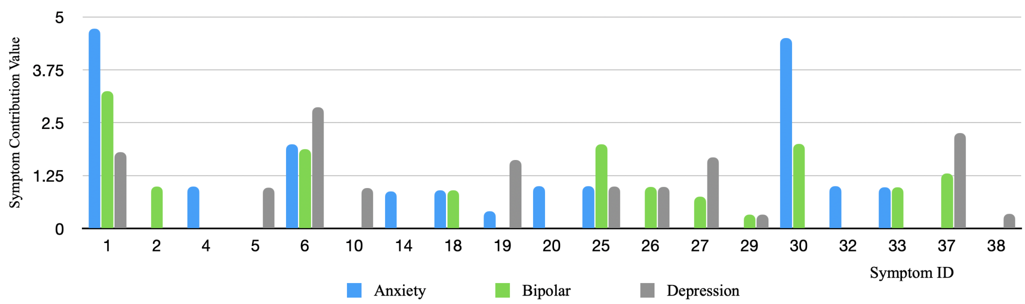

Figure 3: User-level symptom contribution values according to the attention scores obtain from symptom stream model. The user is diagnosed with anxiety, bipolar and depression. We provide the corresponding symptom names with symptom ID in Table 10 (Appendix E), and we omit the symptoms whose contribution values of anxiety, bipolar and depression are all 0.

since there're D disease-specific attention heads. To figure out which symptoms are more critical for diagnosing a certain disease, we measure the *contribution* of each symptom to the detection of different diseases, with the help of attention scores matrix.

For each diagnosed disease d, we calculate the **symptom contribution vector** C_d as follows. First, we select eight symptom probability vectors $S_d = \{s_1, s_2, ..., s_8\}$ with the highest attention score among the symptom stream. Next, for each 38-dimensional vector s_i, we only preserve the value of three symptoms with highest probability. So we set the probability of rest symptoms to 0 and obtain $\hat{s}_i$. Finally, we can get the user's symptom contribution vector

$$C_d = \sum_{i=1}^{8} \hat{s}_i \quad (4)$$

which is demonstrated in Figure 3.

The symptom contribution vectors of these three mental diseases are quite different from each other, satisfying the first aspect of interpretability (i.e., Difference), which means that our proposed model, to some extent, have learned how to "diagnose" a certain disease with its corresponding symptoms just like psychiatrists.

But do these learned high-contribution symptoms truly make sense for the diagnosis? We compare our symptom contribution vectors with the standard criteria DSM-5 to validate the second aspect of interpretability (i.e., Reasonableness). From Figure 3, we can find out that typical symptoms "6: depressed mood"[12] and "37: weight and appetite change" of depression, "1: anxious mood" and "30: panic fear" of anxiety do contribute more to the detection of them than to other diseases, which is in line with DSM-5 . What's more, we can also notice many shared high-contribution symptoms between bipolar disorder and depression, which is reasonable because bipolar disorder contains both depressive episodes and hypomanic episodes[13]. Therefore, we claim that the explicit usage of symptoms, as well as disease-specific attention layers, can exactly improve the interpretability of our neural network model.

3.6 Global Symptom Contribution Analysis

To provide a global view of high-contribution symptoms for all the 7 diseases, we plot a heatmap (See Figure 4) of symptom contribution vectors based on all the users in the test set. Here, we group the users according to his/her diagnosed diseases, and obtain 7 disease-specific user subsets [14]. For each disease subset, we calculate all the user-level symptom contribution vectors as Eq. (4) based on the attention score, and aggregate them by averaging to get the global contribution vector of each disease.

Apart from a lot of agreement reached between our global contribution vector and the authoritative DSM-5 (e.g., "21: drastical shift in mood and energy" for bipolar disorder), we can find several interesting inconsistencies. For example, eating disorder patients are "38: more irritable" and tend to "20: do things that easily get painful consequences", both of which are typical symptoms for bipolar disorder patients in the manic episodes, probably due to the high comorbidity of these two disorders proved by previous researches (Lunde et al., 2009; Ruiz and Gutiérrez-Rojas, 2015). What's more,

[12]We present the symptoms in "ID: name" format.

[13]https://www.nimh.nih.gov/health/topics/bipolar-disorder

[14]a user with multiple mental diseases is in multiple subsets

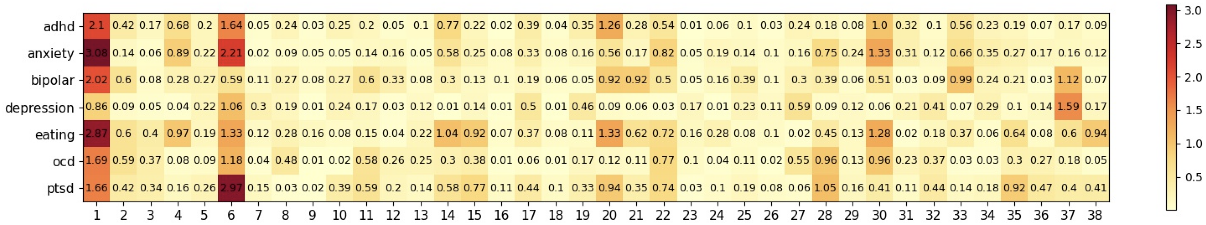

Figure 4: Global symptom contribution vectors of all users in the test set. We provide the corresponding symptom names with symptom id in Table 10, Appendix E.

almost all the diseases treat "1: anxious mood" and "6: depressed mood" as the most important symptoms in PsyEx, while in DSM-5, they're not the typical symptoms for diagnosing many mental diseases such as ADHD. We try to explain this as follows. As a diagnostic criteria, DSM-5 tend to focus on the distinctive symptoms of each disease rather than the common ones. However, these common symptoms can be generally meaningful to distinguish patients of most disorders from control users, so they are highly weighted by our PsyEx model.

4 Related Work

In the literature, substantial efforts have been made to detect a certain mental disease. Some of these works focus on leveraging features like TF-IDF, LIWC (Pennebaker et al., 2001), and posting patterns (Trotzek et al., 2018; Losada and Crestani, 2016) for MDD. Others apply various deep learning methods (Yates et al., 2017; Gui et al., 2019), as well as the contextualized embedding (Ji et al., 2022; Jiang et al., 2020) to improve the performance of classifiers. However, these methods often fail to generalize well (Harrigian et al., 2020) and cannot provide explainable results due to lack of knowledge in the psychiatric domain. To tackle these issues, some works (Lee et al., 2021; Nguyen et al., 2022) began to utilize symptom features, but they extract symptom features with unsupervised/weakly supervised methods, which isn't so reliable for the downstream MDD task.

Recent years, some works start to detect multiple mental disorders. Cohan et al. (2018) proposed a massive Reddit dataset *SMHD* containing 9 mental disorders, followed by many subsequent studies based on this dataset, such as Sekulic and Strube (2019) and Zhang et al. (2022b). However, these works often directly apply a single-disease model to the multi-disease data (i.e., train the model for D times to obtain the results of D diseases), overlooking the correlation among multiple diseases, and thus fail to perform well on rarer diseases.

5 Conclusions

In this work, we tackle the challenge of detecting multiple mental diseases simultaneously in the scenario of comorbidity, and propose a multiple MDD framework achieving SOTA performance in both binary and multi-label setting. We first apply risky post screening based on symptoms, providing reliable diagnostic basis for further disease detection. Then, we propose a two-stream Psychiatric Experts Model with a shared hierarchical encoder and disease-specific attention layers, which simultaneously process the symptom and text features to combine the advantage of both modality. We also explore the interpretability of PsyEx by providing a user-level analysis, as well as measuring the global symptom contribution to the detection of different diseases.

Ethics and Privacy

Though we rely on publicly available Reddit posts in our work, mental health is a sensitive domain and we make efforts to minimize the risk of leaking privacy of individuals in the data collection process. We made no attempt to contact users or link users to other social media accounts. We also replace usernames with random identifiers to prevent users' identities from being known without the use of external information. For the usage of symptom identification dataset, we sign and comply with the data usage agreement to prevent the invasion of privacy or other potential misuses. What's more, the application of mental disease detection should carry out careful analysis and examination. The purpose of this work are not to replace psychiatrists. Instead, we expect our model to be used as an effective auxiliary tool by experienced psychiatrists in the future.

Limitations

Our work has some limitations that could be addressed in future research.

- Though we have made efforts to construct the multiple MDD dataset with a focus on high precision, the disease labels produced automatically by pattern matching of self-reported diagnosis without the guidance of psychiatrists can certainly have some errors. Consequently, our model may inherit the bias from the dataset.

- Despite the significant performance boosting over the baseline, our proposed PsyEx model still cannot achieve satisfying performance in multi-label setting, especially on rarer diseases like eating disorder (See Table 2). To tackle this issue brought by the imbalanced data, we utilize a commonly-used resampling method, which samples equal amount of users with each disease for each batch. However, we find no improvement in the detection performance of these rarer diseases after balanced sampling, indicating that the unsatisfying results aren't just a matter of sparse positive examples. Therefore, we hope to further address this issue in future studies.

- Multiple MDD task is still under-explored currently. Many previous works (e.g., Sekulic and Strube (2019)) only conduct experiments on binary setting (i.e., separately train D models for detecting D mental diseases). Therefore, for the baseline experiments of multi-label setting, we just adopted their hyperparameters on binary setting, which may bring some unfairness.

References

David Adam. 2013. Mental health: On the spectrum. *Nature*, 496:416–418.

DS APA, American Psychiatric Association, et al. 2013. *Diagnostic and statistical manual of mental disorders: DSM-5*, volume 5. American psychiatric association Washington, DC.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018. Smhd: a large-scale resource for exploring online language usage for multiple mental health conditions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1485–1497.

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. From adhd to sad: Analyzing the language of mental health on twitter through self-reported diagnoses. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 1–10.

Caitlin Dreisbach, Theresa A. Koleck, Philip E. Bourne, and Suzanne Bakken. 2019. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. *International Journal of Medical Informatics*, 125:37–46.

Eduardo Fonseca, Shawn Hershey, Manoj Plakal, Daniel PW Ellis, Aren Jansen, and R Channing Moore. 2020. Addressing missing labels in large-scale sound event recognition using a teacher-student framework with loss masking. *IEEE Signal Processing Letters*, 27:1235–1239.

Tao Gui, Liang Zhu, Qi Zhang, Minlong Peng, Xu Zhou, Keyu Ding, and Zhigang Chen. 2019. Cooperative multimodal approach to depression detection in twitter. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*. AAAI Press.

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.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2022. MentalBERT: Publicly available pretrained language models for mental healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7184–7190, Marseille, France. European Language Resources Association.

Zhengping 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, Online. Association for Computational Linguistics.

Femke Lamers, Patricia van Oppen, Hannie C. Comijs, Johannes H. Smit, Philip Spinhoven, Anton J. L. M. van Balkom, Willem A. Nolen, Frans G. Zitman, Aartjan T. F. Beekman, and Brenda Penninx. 2011. Comorbidity patterns of anxiety and depressive disorders in a large cohort study: The netherlands study of depression and anxiety (nesda). *Journal of Clinical Psychiatry*, 72(3):341–348.

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

Shikun Liu, Edward Johns, and Andrew J. Davison. 2019. End-to-end multi-task learning with attention. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1871–1880.

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

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

Anna V. Lunde, Ole B. Fasmer, Kareen K. Akiskal, Hagop S. Akiskal, and Ketil J. Oedegaard. 2009. The relationship of bulimia and anorexia nervosa with bipolar disorder and its temperamental foundations. *Journal of Affective Disorders*, 115(3):309–314.

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

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.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

M. Roca, M. Gili, M. Garcia-Garcia, J. Salva, M. Vives, J. Garcia Campayo, and A. Comas. 2009. Prevalence and comorbidity of common mental disorders in primary care. *Journal of Affective Disorders*, 119(1):52–58.

Eva Ruiz and Luis Gutiérrez-Rojas. 2015. Comorbidity of bipolar disorder and eating disorders. *Revista de psiquiatria y salud mental*, 8(4):232–241.

Ivan Sekulic and Michael Strube. 2019. Adapting deep learning methods for mental health prediction on social media. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 322–327, Hong Kong, China. Association for Computational Linguistics.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2958–2968. Association for Computational Linguistics.

Zhiling Zhang, Siyuan Chen, Mengyue Wu, and Kenny Q. Zhu. 2022a. Psychiatric scale guided risky post screening for early detection of depression. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5220–5226. International Joint Conferences on Artificial Intelligence Organization. AI for Good.

Zhiling Zhang, Siyuan Chen, Mengyue Wu, and Kenny Q. Zhu. 2022b. 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*, page 9970–9985. Association for Computational Linguistics.

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

A Introduction of 7 Mental Disorders

Here, we briefly introduce 7 mental disorders studied in our paper, listing their typical manifestations summarized from DSM-5 in Table 5 for better understanding of these mental disorders.

Disease	Typical Symptoms
ADHD	inattention; hyperactivity and impulsivity.
Anxiety	excessive fear and worry; panic attacks; anxious mood.
Bipolar Disorder	drastic shift in mood and energy; experience periods of mania and depression.
Depression	depressed mood; loss of pleasure or interest; poor concentration; guilty feelings; suicidal ideas.
Eating Disorder	intense fear of gaining weight; binge and purge; rumination; weight and appetite change.
OCD	obsession; compulsion.
PTSD	often develops after a shocking, dangerous event; flashbacks; bad dreams.

Table 5: 7 mental disorders detected in our work with their typical symptoms.

B Detailed Data Statistics

For the multiple MDD dataset (§3.1), we show the number of users suffering from each disease in Table 6. The distribution of the 7 diseases are similar to SMHD (Cohan et al., 2018), and the training/validation/testing set is 8:1:1.

Disease	# Users
Depression	3105
Anxiety	2239
ADHD	2374
Bipolar Disorder	1366
OCD	753
PTSD	391
Eating Disorder	138

Table 6: Number of users suffering from each disease in the MDD dataset.

In addition, the users in this MDD dataset can suffer from multiple mental disorders simultaneously. So we also provide the distribution of the number of diseases on a single user in Table 7. The statistic shows that 57% users suffer from more than one mental disorders, and PsyEx can achieve superior performance for its specific design targeting these comorbidity scenarios.

# Disease	# Users
1	2326
2	1738
3	931
4	407
5	152
6	51
7	14

Table 7: Distribution of a user's disease counts. For example, there are 1738 users suffering from two mental diseases.

C Detailed Experimental Settings of Baselines

For the CNN backbone of *BERT* and *Symp* method, the model structure is the same as that of Nguyen et al. (2022). We train both model with batch size=64, but we set the learning rate as 0.01 when using symptom features, and as 0.003 when using BERT embedding. The *HAN-GRU* model is trained with batch size=32 and learning rate=0.0001. The posting list will be truncated to preserve the earliest 256 posts at most. The average posting number of users in the dataset is 115.2, so 256 is safe enough to preserve almost all the posts of a user.

In the ablation test of different risky post screening methods, the sentence BERT model we utilized for *Similarity* and *K-Means* is `paraphrase-MiniLM-L6-v2`. What's more, the *Similarity* (Zhang et al., 2022a) method extracts key posts according to the cosine similarity between post and mental disease descriptions, which are the description of 38 symptoms (see Table 10) manually summarized from DSM-5.

D Detailed Ablation Results

We show the detailed results of ablation tests in Table 8 and Table 9. Without symptom stream (i.e. w/o symp-stream) or disease-specific attention layer (i.e. w/o multi-attn), the F1 score on nearly all the diseases dropped, especially the rarer diseases like eating disorder and OCD. What's more, we can notice a significant drop in the performance without symptom-based risky post screening in all the diseases, suggesting the importance of a precise screening method to filter out the noisy data.

D.1 Impact of the Selected Posts Number

Here we study the impact of the number of posts selected in risky posts screening (see Figure. 5). We can observe that 16 posts have the best performance for nearly all the diseases. To find out the reason,

Model	Depression	Anxiety	ADHD	Bipolar	OCD	PTSD	Eating	Avg. F1
Two-Stream PsyEx	87.89	89.84	84.4	91.58	81.69	81.69	85.71	**86.12**
w/o symp-stream	88.12	89.09	84.55	89.33	76.71	81.58	83.33	84.67
w/o multi-attn	87.64	88.64	82.84	89.7	76.39	79.49	81.08	83.68
w/o multi-task	87.69	88.89	84.21	91.16	75	84.21	86.67	85.40

Table 8: Ablation tests for the design choices of PsyEx, reporting F1 score of each disease (detailed results of Table 3).

Screen method	Depression	Anxiety	ADHD	Bipolar	OCD	PTSD	Eating	Avg. F1
Symptom-based (PsyEx)	87.89	89.84	84.4	91.58	81.69	81.69	85.71	**86.12**
Similarity	86.84	87.7	83.03	87.84	78.26	77.14	76.47	82.47
K-Means	79.34	81.15	68.66	82.0	70.34	73.42	68.42	74.76
Last	69.27	67.89	55.31	67.6	45.8	46.15	45.16	56.74

Table 9: Ablation tests for different risky post screening methods, reporting F1 score of each disease (detailed results of Table 4)

we calculate the average symptom probability of the selected K posts sorted by its highest symptom probability, which is 0.25 for the posts ranked 16, meaning that 16 posts is enough to include most of the symptomatic posts and adding more posts can easily introduce some noisy data.

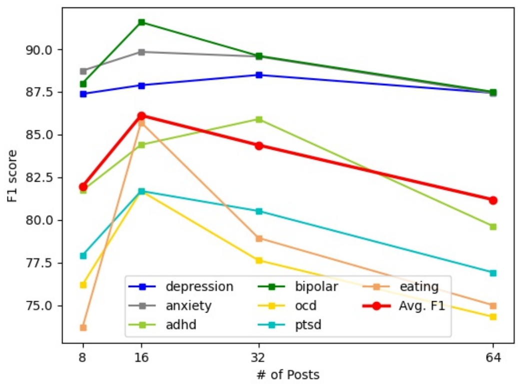

Figure 5: Impact of the number of selected posts on each disease and the mean F1.

E Psychiatric Symptoms

We use serial numbers to represent symptoms in Figure 3 and Figure 4, so we provide the corresponding symptom names in Table 10 for reference.

id	Symptom
1	Anxious Mood
2	Autonomic symptoms
3	Cardiovascular symptoms
4	Catatonic behavior
5	Decreased energy tiredness fatigue
6	Depressed Mood
7	Gastrointestinal symptoms
8	Genitourinary symptoms
9	Hyperactivity agitation
10	Impulsivity
11	Inattention
12	Indecisiveness
13	Respiratory symptoms
14	Suicidal ideas
15	Worthlessness and guilty
16	Avoidance of stimuli
17	Compensatory behaviors to prevent weight gain
18	Compulsions
19	Diminished emotional expression
20	Do things easily get painful consequences
21	Drastic shift in mood and energy
22	Fear about social situations
23	Fear of gaining weight
24	Fears of being negatively evaluated
25	Flight of ideas
26	Intrusion symptoms
27	Loss of interest or motivation
28	More talkative
29	Obsession
30	Panic fear
31	Pessimism
32	Poor memory
33	Sleep disturbance
34	Somatic muscle
35	Somatic symptoms others
36	Somatic symptoms sensory
37	Weight and appetite change
38	Anger Irritability

Table 10: Id and its corresponding symptoms