From Detection to Discovery: A Closed-Loop Approach for Continuous Medical Knowledge Expansion and Depression Detection on Social Media

Online Supplementary Material

Appendix 1: Initial Knowledge graph Evaluation

To construct the depression diagnosis-related knowledge graph, we have drawn from widely recognized medical literature and depression clinical screening measures, including:

1. APA. DSM-5-TR(tm) Classification. American Psychiatric Association Publishing, 2022.

2. Beck, M.D.A.T. and Alford, P.D.B.A.Depression: Causes and Treatment. University of Pennsylvania Press, 2014.

3. Martin, A., Rief, W., Klaiberg, A., and Braehler, E. Validity of the Brief Patient Health Questionnaire Mood Scale (PHQ-9) in the general population. General Hospital Psychiatry, 28, 1 (January 2006), 71–77.

4. Rush, A.J., Trivedi, M.H., Ibrahim, H.M., et al. The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), Clinician Rating (QIDS-C), and Self-report (QIDS-SR): A Psychometric Evaluation in Patients with Chronic Major Depression. Biological Psychiatry, 54, 5 (September 2003), 573–583.

To validate the constructed depression diagnosis-related knowledge graph, we employ a two-step validation process. The first step involves a medical knowledge coverage evaluation. The second step includes a medical expert evaluation. In the first step, the evaluation of medical knowledge coverage aims to involve assessing the ontology's coverage by comparing the number of concepts within the ontology to several commonly utilized depression diagnostic scales. The coverage results of the depression ontology with respect to DSM-5-TR, PHQ-9, and QIDS-SR are detailed in Table A1. The coverage calculations demonstrate that our ontology comprehensively encompasses the content of widely used depression scales.

Table A1. The Coverage Rate of Depression Ontology to Depression Diagnosis Scale

Depression diagnosis scales	**DSM-5-TR**	**PHQ-9**	**QIDS-SR**
Coverage rate	85.6%	95.2%	93.8%

In the second step of evaluation: medical expert assessment, we engaged a panel of medical experts. Two psychiatrists from a nationally recognized hospital were invited to evaluate our depression ontology. Following their initial independent reviews, the psychiatrists convened to include entities that were absent in the initial version while eliminating redundant and clinically irrelevant entities. Entity classes that were misassigned were corrected. The final ontology, approved by these medical experts, was employed in our study.

Appendix 2: Parameter Settings, Testing Data, and Knowledge Graph Pre-training

For the implementation of our method, we use PyTorch to build the proposed method. The embedding dimension of Relational Graph Neural Network with Hierarchical Attention (RGHAT) is set to 768, while that of the large language model (LLM) is set to 4096. The maximum sequence length is set to 100 for the depression-related entity recognition component and 200 for the depression detection module. For the other deep learning-based baseline algorithms, the maximum sequence length is set to 200. All deep learning benchmarks adopt the AdamW optimizer for parameter optimization, and the learning rate is set to 1e-4.

Table A2. Parameter Settings for Compared Methods

Model	**Parameter**	**Value**	**Model**	**Parameter**	**Value**
Choudhury et al. (2013)	Regularization parameter	1	LSTM-based (Khan et al., 2021)	Number of hidden units	256
	Kernel	Linear	LSTM-based (Khan et al., 2021)	Number of hidden units	500
Coppersmith et al. (2014)	Regularization parameter	1	Attention-based (Ghosh et al., 2023)	Number of hidden units of LSTM	128
	Solver	L-BFGS		Number of hidden units of CNN	32

Preotiuc-Pietro et al. (2015)	Regularization parameter	1	Attention-based (Thekkekara et al., 2024)	Number of hidden units of LSTM	128
	Solver	L-BFGS		Number of hidden units of CNN	128
Benton et al. (2017)	Number of hidden layers	2	Contrastive Learning-based (Liu et al., 2024)	Dynamic routing num	3
	Number of hidden units	[20, 10]		Number of hidden units	768
Reece et al. (2017)	Criterion	entropy	Transformer-based (Malviya et al., 2021)	Number of heads	12
	Number of estimators	750		Number of layers	10
				Number of hidden units	768
Chau et al. (2020)	Estimators	SVC + RF + MLP	JDeC (Ours)	Number of hidden units of RGHAT	768
CNN-based (Lin et al., 2020)	Conv 1/2/3 kernel size	[3, 4, 5]		Number of hidden units of LLM	4096
	Number of kernels	256		LoRA rank	4
	Pool	Max pool		LoRA alpha	8

The statistics of the testing dataset has been provided in Table A3.

Table A3. Dataset

	Depressed	Non-depressed
Number of posts	141,736	1,724,788
Number of words	4,244,173	48,002,886
Average number of posts per user	310	505
Average number of days from first to last post	620	621
Number of users	457	3413

The knowledge graph pre-training is detailed below. In the pre-training stage, the knowledge graph learns embeddings from labeled social media posts generated by clinically diagnosed depression patients on WebMD, a widely recognized online health information platform. On WebMD, patients discuss their depression symptoms, treatments, and life events, which are depression-related entities. This provides a foundation for modeling the social media behavior of depression patients. In a dataset previously used in existing research, depression-related entities in WebMD posts were manually labeled and validated by experts (Zhang et al. 2024). Additional labeling was conducted by inviting mental health professionals to classify depression-related entities into psychological symptoms and physical symptoms, and to further classify treatment-related entities into medication and therapy. Entities that co-occur within the same post are paired and connected using predefined relations, forming a set of triplets that serve as positive samples for entity embedding learning (as described in Section 3.1.2).

Appendix 3: The Algorithm for Entity Importance Estimation.

The algorithm for estimating entity importance in our proposed framework is presented in Algorithm A1.

Algorithm A1. Entity Importance Computation

Input: Depression knowledge graph G and an entity e_i
Output: The max transition probability p_{trans} as the importance value of e_i for depression prediction
Set e_i as the start node and $e_{depression}$ as the target node
Set the maximum transition path length as p_{length}
Set the number of Monte Carlo samples as $n_{samples}$
Procedure MonteCarlo_MaxTransProb(G, e_i, $e_{depression}$, p_{length}, $n_{samples}$):
    p_{trans} = 0
    For $i \leftarrow 1, \ldots, n_{samples}$ **do**
        current_node ← e_i
        $path\_probability$ ← 1.0
        $path\_length$ ← 0
        While not (current_node==$e_{depression}$ or $path\_length > p_{length}$) **do**
            $N_{current}$ ← the neighbor entities for current_node in G
            If $N_{current}$ is empty **then**
                End while
            next_node ← randomly choose an entity from $N_{current}$
            $edge\_weight$ ←transition score r from current_node to next_node
            $path\_probability = path\_probability \times edge\_weight$
            current_node ← next_node
            $path\_length = path\_length + 1$
        If current_node == $e_{depression}$ and $path\_length \le p_{length}$ **then**
            $p_{trans} = max(p_{trans}, path\_probability)$
    Return p_{trans}