HOPE: Hybrid Optimized Parallel Encoding with Supervised and Unsupervised Semantic Fusion for Depression Symptom Detection

Tu-Phuong Mai[1*], **Minh-Ha H. Le**[1*], **Duc-Luong Tran**[1], **Phuong-Anh Chu**[1], **Duy-Cat Can**[1,2,3†], and **Hoang-Quynh Le**[1†]
[1]Faculty of Information Technology, VNU University of Engineering and Technology, Vietnam
[2]Platform of Bioinformatics, Lausanne University Hospital, Switzerland
[3]Faculty of Biology and Medicine, University of Lausanne, Switzerland
{21020552, 21020621, 22021148, 23020324, lhquynh}@vnu.edu.vn
duy-cat.can@chuv.ch

Abstract

Timely detection of depression symptoms is essential for early intervention, and the continuous stream of user-generated content on social media provides an ideal source for this purpose. To address this challenge, we propose HOPE, a **H**ybrid **O**ptimized **P**arallel **E**ncoding framework that combines supervised symptom relevance signals with unsupervised intrinsic semantic clustering. This parallel design enables robust symptom detection under limited labeled data and introduces a distinctive semantic-similarity perspective with automatic class-anchor adjustment. We also propose an optimized hybrid semantic fusion mechanism to combine supervised and unsupervised scores through a learnable module. We evaluate our system on multiple benchmark datasets and surpass previous approaches, demonstrating its effectiveness in detecting fine-grained symptoms and early warning of mental health risk. Source code is available at https://github.com/candleMind/hope.

1 Introduction

As one of the leading causes of disability worldwide, depression is increasingly projected to become the primary contributor to the global disease burden by 2030 (Zhang et al., 2024). Despite its prevalence, it remains widely underdiagnosed due to social stigma and limited access to mental health screening. Alarmingly, nearly 60% of young people who died from suicide had shown signs of depression without ever receiving a clinical diagnosis (Chaudhary et al., 2024), highlighting the urgent need for early and sensitive detection methods. In this context, social media platforms have emerged as a promising source for detecting early signs of depression, because they provide unfiltered access to users' thoughts and emotions in real time.

This work addresses the problem of detecting depressive symptoms from social media data through multi-label classification and ranking. We focus on symptom-level detection because it provides clinically grounded evidence aligned with diagnostic frameworks (e.g., BDI-II, DSM-5), enhances model reliability and provides more insights into psychological distress. Figure 1 illustrates the challenges of this problem with a Reddit post containing depressive symptoms, showing implicit psychological distress and the semantic overlap among related symptoms.

Figure 1: A mock example of sentence-level multi-label annotations from a Reddit post.

Beyond the task challenges, a major limitation is the scarcity and subjectivity of labeled data, especially for multi-label tasks, where limited labels can lead to biased learning (Ma and Chen, 2021). Supervised models also struggle with the temporal change in data properties, or temporal domain shift between training and testing data, due to changes in users' topic distributions and complex psychological dynamics (Guo et al., 2022). Meanwhile, abundant unlabeled data remains largely underexploited (Wu et al., 2024), while unsupervised models also struggle with multi-label instances positioned between class boundaries (see Section 5.1). This calls for a method that exploits the complementarity of supervised and unsupervised modeling. Furthermore, integrating domain knowledge

[*] Co-first authors.
[†] Corresponding authors.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2898–2911

into modeling enhances model reliability by guiding the learning process with expert insights and prior information (Wang et al., 2021).

In response to these challenges, our primary contributions are threefold: **(1) We propose a novel framework HOPE (Hybrid Optimized Parallel Encoding)** to combine *supervised* depressive pattern recognition from labeled data with *unsupervised* semantic clustering applied directly to the unlabeled testing set (transductive setting). By treating the test corpus (e.g., the 17M eRisk texts) as unlabeled data to simulate real-world inference, HOPE effectively tackles label scarcity by exploiting intrinsic characteristics of the target domain. We also propose an *optimized hybrid semantic fusion* mechanism to dynamically leverage the strengths of both paradigms through a learnable module. **(2) Intrinsic semantic clustering with domain knowledge.** The framework integrates domain knowledge by initializing clustering centroids with few-shot-enhanced semantic descriptions of target labels, such as clinically defined symptoms or diagnostic criteria. This unsupervised modeling introduces a distinguished semantic-similarity perspective to the supervised classification by automatic class-anchor adjustment based on the intrinsic structure of the test data. **(3) Evaluation on multiple benchmark social-media datasets.** Our framework shows robust performance on both multi-label relevance ranking and classification tasks across datasets.

2 Related Work

Depression Classification. Depression detection on social media has been formulated at both the post and user levels (Ta et al., 2025). Early approaches combined sentiment features with topic modeling (LDA) (Zogan et al., 2021), while Tejaswini et al. (2024) proposed hybrid models integrating FastText, CNN, and LSTM for enhanced text representation. The emergence of transformer models marked a significant advance, with Helmy et al. (2024) leveraging BERT to capture nuanced contextual signals from user text. Recent work has shifted toward multimodal and temporal modeling to capture the dynamic nature of psychological states (Bucur et al., 2023; Nguyen et al., 2025).

Depressive Symptom Detection. To address the limited explainability in binary depression classification, recent research has shifted toward symptom-level detection for clinically grounded, interpretable predictions (Bao et al., 2025). Milintsevich et al. (2024) marked input text with sentiment, emotion, and domain-specific lexicon terms to guide transformer-based PHQ-8 symptom prediction. Lee et al. (2024) combined KLUE-RoBERTa with explainable AI (SHAP) to identify 12 symptoms from emergency dialogues. Large language models (LLMs) have also been explored for symptom extraction and explanation: Wang et al. (2024) used LLM encoders with similarity measures for symptom evidence extraction, while Bao et al. (2024) employed LLMs for few-shot explanation generation. However, the fidelity of LLM-generated explanations remains difficult to quantify. Supporting datasets include PsySym (Zhang et al., 2022) for multi-disease symptoms, PRIMATE (Gupta et al., 2022) for PHQ-9 annotations, ReDSM5 (Bao et al., 2025) for DSM-5-aligned symptoms, and DepressionEmo (Rahman et al., 2024) for depressive emotion annotations.

Ranking-based approaches prioritize sentences by relevance to specific symptoms, thereby improving diagnostic utility and confidence levels. APB-UC3M (Bascuñana and Bedmar, 2024) and DS@GT (Guecha et al., 2024) employed binary classifiers with RoBERTa or BDI-II questions (Beck et al., 1961) but suffered from poor calibration. The BLUE team (Parapar et al., 2023) used ChatGPT-generated queries but remained dependent on predefined structures. However, in low-resource settings (Parapar et al., 2024), these supervised methods struggle because they require large amounts of labeled data, poor generalization with domain shifts, and inability to exploit unlabeled testing data.

Unsupervised and Hybrid Learning. Unsupervised learning identifies inherent data structures without labels. In psychiatry, such algorithms have discovered depression subtypes, transcending existing diagnostic labels (Squires et al., 2023). Neural embeddings, particularly in-domain Word2Vec and Doc2Vec, substantially outperform TF-IDF for clustering-based classification (Kosar et al., 2022). An et al. (2025) combined pretrained features with UMAP and hybrid clustering voting for improved classification. Similarity-based approaches face generalization challenges. While similarity methods with SimCSE or SBERT embeddings outperform zero-shot approaches (Schopf et al., 2022), defining class anchors remains difficult, and LLM-based similarity methods often strug-

gle across diverse symptom expressions (Wang et al., 2024). Hybrid methods combining unsupervised and supervised paradigms show promise: Yu et al. (2025) used spectral clustering for pseudo-labeling followed by semi-supervised SVMs, while Alrammahi et al. (2025) leveraged clustering for pseudo-labels then trained supervised models.

Research Gaps. Despite these advances, critical gaps remain. Existing hybrid supervised-unsupervised methods fail to simultaneously exploit both training and testing data, typically using unsupervised learning only for pseudo-labeling rather than parallel inference. In addition, integrating domain knowledge into unsupervised models by leveraging clinical symptom descriptions to identify plausible class anchors is still largely unexplored.

3 Methodology

3.1 Overview of Proposed Model

We formulate depressive symptom detection as a multi-label problem. Given a set of candidate texts $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$ extracted from user-generated social media posts, and a predefined set of depression symptoms $\mathfrak{L} = \{l_1, l_2, \ldots, l_n\}$, our objective is to learn a function $f : \mathcal{S} \times \mathfrak{L} \to \mathbb{R}$ that assigns relevance scores to each text-symptom pair (s_i, l_j), where higher scores indicate stronger association between text s_i and symptom l_j.

HOPE fills the aforementioned research gaps through a three-component architecture, as described in Figure 2: (1) a *Supervised Symptom Extractor* that leverages DepRoBERTa to capture depressive patterns from labeled training data and generate initial relevance scores; (2) an *Unsupervised Intrinsic Semantic Clustering* module that employs few-shot-enhanced symptom-initialized K-means clustering, with paired-stream BERT embeddings, to extract intrinsic semantic relationships directly from unlabeled testing inputs; and (3) an *Optimized Hybrid Semantic Fusion* mechanism that leverages the dual views of supervised and unsupervised modeling via a learnable Multilayer Perceptron (MLP) module.

3.2 Supervised Symptom Extractor

This module addresses the challenge of learning depressive patterns from limited labeled training data by leveraging domain-specific pre-trained representations. The module takes a text s_i as input and outputs symptom relevance scores $\mathbf{r}_i = [r_{i,j}]_{j=1}^{|\mathfrak{L}|}$,

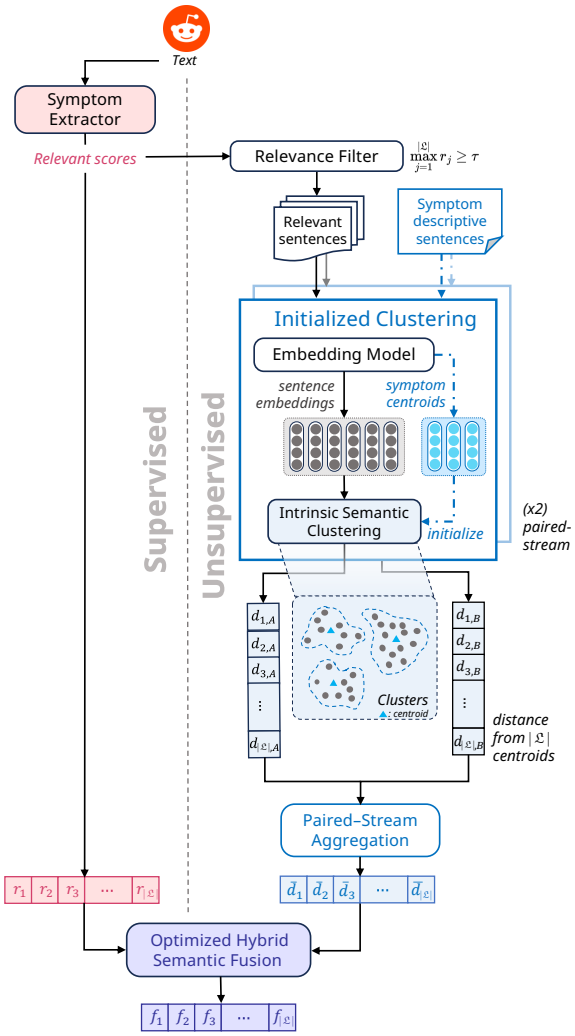

Figure 2: Overview of the proposed Hybrid Optimized Parallel Encoding for detecting depression symptoms.

where $r_{i,j}$ denotes the relevance of text s_i to symptom l_j, and $|\mathfrak{L}|$ is the total number of symptoms considered.

We employ DepRoBERTa (Poświata and Perełkiewicz, 2022) as the backbone model, which is pre-trained on depression-specific social media data and enables effective transfer of depressive linguistic patterns to fine-grained symptom detection. We adapt the model through fine-tuning for multi-label classification with a modified architecture that incorporates symptom-specific classification heads. The supervised extraction process is formulated as:

$$\mathbf{h}_i = \text{DepRoBERTa}(s_i) \quad (1)$$

$$\mathbf{r}_i = \sigma(\mathbf{h}_i \mathbf{W}_{\text{cls}} + \mathbf{b}_{\text{cls}}) \quad (2)$$

where $\mathbf{h}_i$ is the contextualized representation of sentence s_i, σ is the sigmoid function, $\mathbf{W}_{\text{cls}} \in \mathbb{R}^{\dim_{\mathbf{h}} \times |\mathfrak{L}|}$ and $\mathbf{b}_{\text{cls}} \in \mathbb{R}^{|\mathfrak{L}|}$ are learnable parameters.

3.3 Unsupervised Intrinsic Semantic Clustering

While effectively extracting depressive features from labeled data, supervised models often struggle with the domain shift between training and testing distributions. To address this limitation, this module introduces a novel approach that directly leverage intrinsic information from unlabeled testing data through *few-shot-enhanced, symptom-initialized clustering*. In addition, by incorporating unsupervised learning, we augment the traditional supervised patterns with a distinctive semantic-similarity perspective to measure class relevance.

In contrast to existing approaches that rely on direct similarity calculations between embeddings of items and class descriptions, we employ a clustering algorithm for automatic class-anchor adjustment to the target domain. By using clinical symptom descriptors as initial clustering centroids, the algorithm enables the resulting centroids to capture the specific features of each class within the unlabeled testing data. Furthermore, the proposed paired-stream embedding strategy mitigates individual model biases and enhances semantic coverage, thereby ensuring robust clustering.

The module processes texts filtered by the supervised relevance scores and outputs unsupervised similarity scores $\mathbf{d}_i$ of each text s_i to each symptom l_j. Clustering is performed at the sentence level for better precision, thus, for post-level texts, $\mathbf{d}_i$ is obtained by taking the maximum value across sentences in the post.

Relevance Filter. Our experiments show that direct clustering struggles with scattered and unpredictable patterns of the non-relevant class (see Section 5.1). To address this problem, sentences with minimal depressive content are removed to reduce noise and enhance cluster discriminability, ensuring that subsequent clustering focuses on symptom-relevant semantic patterns. Specifically, we discard sentences where all symptom relevance scores fall below the threshold $\tau = 0.1$. This aggressive filtering removes over 95% of the 17 million sentences in the eRisk dataset while prioritizing high recall for symptom-bearing content:

$$\mathcal{S}_{\text{filtered}} = \{s_i \in \mathcal{S} : \max_{j=1}^{|\mathfrak{L}|} r_{i,j} \geq \tau\} \tag{3}$$

Symptom-Initialized K-means Clustering. This module aims to extract symptom-specific semantic relationships among sentences while addressing the limited generalizability of symptom description anchors on unseen test data. To this end, we apply K-means clustering with centroids initialized as the averaged embeddings of each symptom's descriptive sentences, allowing for automatic class-anchor adjustment. We set the number of clusters K equal to the number of symptoms $|\mathfrak{L}|$, then iteratively update these centroids on the unlabeled test data. Given the importance of centroid initialization in K-means, we leverage few-shot prompting on LLMs to generate symptom descriptions more closely aligned with real-world data. The few-shot-enhanced symptom descriptions are generated by `Claude Sonnet 4`, integrating clinical knowledge from diagnostic frameworks (e.g., BDI-II, DSM-5), and 20 randomly selected seed samples from the training set. The outputs include five different sentences for each symptom, varying in expression and severity. Specifically, for each symptom l_j, we compute the initial centroid $\mathbf{c}_j$ as:

$$\mathbf{c}_j = \frac{1}{|\mathcal{Q}_j|} \sum_{q \in \mathcal{Q}_j} \mathcal{E}(q) \tag{4}$$

where $\mathcal{Q}_j$ represents the set of enhanced descriptive sentences for symptom l_j (see Appendix C), and $\mathcal{E}(q)$ denotes the embedding of sentence q, derived from the below embedding models.

Paired-stream Embeddings. This paired-stream architecture uses two embedding models to mitigate individual model biases, enhance semantic coverage, and capture different aspects of semantic similarity, which is crucial for robust clustering. We employ two complementary embedding models: `nomic-embed-text-v1.5`[1] (denoted as model $\mathcal{E}_A$) and `modernbert-embed-base`[2] (denoted as model $\mathcal{E}_B$). Both models are trained using contrastive learning with vector-distance metrics to learn sentence-level semantic embeddings. Consequently, they are better optimized for distance-based measures compared to the DepRoBERTa embeddings (in Section 3.2), where features may not be equally representative, potentially introducing noise into the embedding distance.

For each embedding model $\mathcal{E}_k$ ($k \in \{A, B\}$), we perform K-means clustering:

[1] https://huggingface.co/nomic-ai/nomic-embed-text-v1.5

[2] https://huggingface.co/nomic-ai/modernbert-embed-base

$$d_{i,j,k} = \frac{1}{1 + \|\mathcal{E}_k(s_i) - \mathbf{c}_{j,k}\|_2} \quad (5)$$

where $\mathcal{E}_k(s_i)$ is the embedding of sentence s_i from model $\mathcal{E}_k$, and $\mathbf{c}_{j,k}$ is the centroid for symptom l_j in embedding space k.

The final unsupervised scores are computed by averaging the outputs of both embedding models:

$$\mathbf{d}_i = [\bar{d}_{i,j}]_{j=1}^{|\mathfrak{L}|} \quad (6)$$

$$\bar{d}_{i,j} = \frac{1}{2}(d_{i,j,A} + d_{i,j,B}) \quad (7)$$

3.4 Optimized Hybrid Semantic Fusion

This module introduces a learnable hybrid mechanism that combines semantic information from supervised relevance scores $\mathbf{r}_i$ and unsupervised similarity scores $\mathbf{d}_i$ to produce final scores $\mathbf{f}_i = [f_{i,j}]_{j=1}^{|\mathfrak{L}|}$, instead of common rule-based late fusion such as averaging or maximizing. This mechanism addresses the complementary strengths and limitations of dual views: supervised models excel at extracting depressive features from labeled data but may suffer from domain shift between training and testing data, while unsupervised models capture intrinsic semantic relationships in unlabeled data but struggle with multi-label instances positioned between class boundaries. We introduce a trainable MLP module that automatically learns optimal symptom-view weights per dataset while facilitating information exchange among related symptoms:

$$\mathbf{f}_i = \text{MLP}(\mathbf{r}_i \oplus \mathbf{d}_i) \quad (8)$$

where $\oplus$ denotes concatenation. During training, the model uses r_i and d_i from the training set as inputs and the labels of s_i as targets, and then applies the trained module on the testing set.

For each symptom l_j, the score $f_{i,j}$ is used to select the top 1000 highest-scoring texts for the ranking, or to classify with a pre-defined threshold.

4 Experimental Results

4.1 Experimental Setup

4.1.1 Tasks and Datasets

We evaluate our approach on two tasks using four English-language Reddit datasets, all of which are anonymized to protect user privacy.

Symptom Relevance Ranking. We follow the official benchmark for Task 1 of the CLEF eRisk 2025 (Parapar et al., 2025b) shared task, which focuses on the *Search for Symptoms of Depression* task. For each of the 21 BDI-II symptoms (Beck et al., 1961), participants must retrieve and rank up to 1,000 relevant sentences from a massive Reddit corpus of over 17 million sentences. The training data, containing more than 19 million sentences but only approximately 28,000 labeled instances, is derived from the eRisk 2023 and 2024 shared tasks (Parapar et al., 2023, 2024).

Symptom Classification. We additionally evaluate our model on three multi-label depression symptom datasets that are derived from Reddit:

- *ReDSM5* (Bao et al., 2025): Consists of 1,896 sentences from 1,484 posts, re-annotated by a licensed psychologist following nine *DSM-5* symptoms (Vahia, 2013).
- *PRIMATE* (Gupta et al., 2022): Comprises 2,003 long-form posts, annotated by five crowd workers under expert supervision based on the nine *PHQ-9* symptoms (Kroenke et al., 2001).
- *DepressionEmo* (Rahman et al., 2024): Includes 6,037 posts annotated with eight emotional symptoms via weak supervision (majority voting across four transformer-based classifiers), with reliability verified on random samples.

Detailed statistics and additional dataset-level information are summarized in Table 1.

4.1.2 Evaluation Metrics

We employ different evaluation metrics for the two tasks in our framework.

Ranking Task. Following the official eRisk 2025 benchmark, we use four standard ranking-based metrics: (i) *Mean Average Precision* (`MAP`), representing the mean of the precision values at the ranks where relevant sentences appear; (ii) *R-Precision* (`R-PREC`), defined as precision at the rank equal to the total number of relevant sentences for a given symptom; (iii) *Precision at 10* (`P@10`), the fraction of relevant sentences among the top-10 retrieved results; and (iv) *Normalized Discounted Cumulative Gain at 1000* (`NDCG`), which accounts for the ranking positions of relevant sentences and rewards highly ranked correct results.

Datasets	Ranking	Classification		
	eRisk 2025 (Train / Test)	ReDSM5	PRIMATE	DepressionEmo
Total texts	19,349,315 / 17,553,441	1,896	2,003	6,037
Labeled texts	28,018 / 10,383	1,896	2,003	6,037
Avg. words	17.12 / 12.39	13.81	246.56	103.97
Train / Val / Test	25,216 / 2,802 / 10,383	1,288 / 228 / 380	1,361 / 241 / 401	4,225 / 906 / 906
Annotation unit	Sentence	Sentence	Post	Post
Annotation type	Human annotation	Human annotation	Human annotation	Weak supervision
Labels	21 symptoms (BDI-II)	9 symptoms (DSM-5)	9 symptoms (PHQ-9)	8 symptoms

Table 1: Corpus statistics across datasets. eRisk 2025 is used for the ranking task, while ReDSM5, PRIMATE, and DepressionEmo are utilized for classification tasks. For eRisk, we report training and testing sets separately.

Classification Task. We adopt *Macro-Precision* (Mac-P), *Macro-Recall* (Mac-R), *Macro-F1* (Mac-F1), and *Micro-F1* (Mic-F1) as evaluation metrics. Macro scores assess class-balanced performance, while *Micro-F1* reflects overall accuracy across all instances.

4.1.3 Comparative models

We evaluate our approach in two settings: ranking and classification tasks.

Ranking Task. We compare our approach with the top four teams in Task 1 of eRisk 2025 (Parapar et al., 2025a). *INESC-ID* (Nunes and Ribeiro, 2025) utilizes an ensemble of fine-tuned DeBERTa models with LLM-based data augmentation and semantic similarity scoring against BDI-II items. *UET-Psyche-Warriors* (Mai et al., 2025) combines two approaches: K-means clustering to derive symptom prototypes with KNN label assignment, and a multi-task fine-tuned RoBERTa model with contrastive learning for joint symptom and severity detection. However, they rely on supervised learning from training data, without addressing the domain shift problem. *SonUIT* (Son and Thin, 2025) applies a two-stage pipeline using symptom embedding similarity followed by re-ranking with cross-encoders and BM25. *PJs-team* (Vachharajani, 2025) developed bi-encoder and cross-encoder systems, which combine fine-tuning, ensembling, and reranking strategies, for query-based prediction using BDI-II items.

Classification Task. We evaluate our model against diverse models across three benchmark datasets, spanning traditional classifiers such as *SVM*, *LightGBM*, and *XGBoost* (Bao et al., 2025; Rahman et al., 2024); deep learning models including *CNN*, and *Stacked LSTM* (Bao et al., 2025; Younas et al., 2025); and transformer-based methods such as *BERT*, *RoBERTa*, *BART*, and *LLM (LLaMA-3.2-1B)* (Zhang et al., 2023; Rahman et al., 2024; Violides et al., 2024; Bao et al., 2025). We also include specialized variants: *MentalBERT* pretrained on mental health corpora, *FL-BERT* augmenting BERT with label embeddings (Zhang et al., 2023), *GAN-BERT* with adversarial training (Rahman et al., 2024), and *SelfAug-CL* using self-supervised contrastive learning (Khan et al., 2024). Methods specifically designed for multi-label classification include *SpanEmo* for span-based emotion detection, *LEAR* enhancing spans with label knowledge, *aMLP* improving attention mechanisms, *LR-GCN* integrating label graphs via GNNs, and *Span-PHQ* aligning posts with PHQ-9 descriptions via span prediction and contrastive learning (Zhang et al., 2023). Other explainable models include *PSAT* with PHQ-9-infused cross-attention (Dalal et al., 2025) and *KiNN2* leveraging multi-level knowledge infusion (Dalal et al., 2024).

4.2 Overall Results and Analysis

Ranking Task. Table 2 reports the ranking performance of our proposed HOPE framework compared with the top four highest-performing teams submitted to the official eRisk 2025. Across three out of the four ranking-based metrics, HOPE achieves the highest scores. Specifically, compared to the top-performing team *INESC-ID (unanimity strategy)*, our approach outperforms their results by an improvement of 4.05% in MAP, 2.54% in R-PREC, and 11.79% in NDCG, while maintaining a competitive P@10 score. The improvement in NDCG is particularly notable, with a 3.99% gain over *INESC-ID's best* NDCG *score (max strategy)*, highlighting our model's ability to rank relevant sentences with high confidence.

Classification Task. As shown in Table 3, HOPE outperforms all compared methods across the three

Team / Submission	MAP	R-PREC	P@10	NDCG
INESC-ID *(Rank 1)* †				
max	35.00	40.70	64.80	65.30
unanimity	35.40	43.30	**87.60**	57.50
UET *(Rank 2)* †				
ensemble similarity	31.50	39.00	65.70	61.20
machine learning	33.90	39.40	77.60	62.30
SonUIT *(Rank 3)* †				
config 4	32.80	42.60	76.70	57.80
PJs-team *(Rank 4)* †				
teamRRens-v2	27.90	36.00	80.00	50.30
HOPE (Our) ‡	**39.45** ±1.00	**45.84** ±0.91	81.43 ±0.72	**69.29** ±0.88

*Best scores in **bold**, second best underlined.*
†*: Competition official results (Parapar et al., 2025a).*
‡*: Our results are mean ± standard deviation over 10 runs.*

Table 2: Ranking-based evaluation results for our method and the top 4 teams on Task 1 eRisk 2025 (%).

datasets. On *ReDSM5*, it improves upon the strongest baseline (fine-tuned LLM *LLaMA-3.2-1B*) by 27.15% Macro-F1 and 23.87% Micro-F1. On *PRIMATE*, it surpasses advanced approaches, including span-based and label-aware models, with gains of 1.79% Macro-F1 and 1.82% Micro-F1 over the state-of-the-art *SpanPHQ*. On *DepressionEmo*, HOPE exceeds *SelfAug-CL* by 1.53% Macro-F1 and outperforms fine-tuned *RoBERTa* by 1.81% Micro-F1. These results demonstrate the robustness and generalizability of our framework across diverse data distributions.

4.3 Model Component Contribution Analysis

We investigate the contribution of each component to system performance by conducting ablation experiments on both tasks, ranking (on the eRisk dataset) and classification (on the ReDSM5 dataset), by ablating each component in turn and analyzing the resulting impact, as shown in Figure 3. Both the *Supervised Symptom Extractor* and the *Unsupervised Intrinsic Semantic Clustering Branch* are the most critical components. Removing either component leads to substantial performance degradation: without supervision, the model struggles to capture the broad spectrum of depressive symptoms; conversely, omitting the unsupervised branch results in a significant decline due to the 1–2-year gap (domain shift) between the eRisk training and testing sets. Our *Centroid Initialization and Semantic Clustering Strategy* is key to stable and domain-adaptive clustering. Performance drops significantly when using random initialization (*W/o Centroid Init*), initialization using original clinical symptom descriptions (*W/o LLM-Init Centroid*), or training-data-based centroids (*W/o Intrinsic Clustering*). Disabling the *Relevance Filter* forces the model to cluster all candidate sentences, including irrelevant ones, thereby introducing noise and lowering performance. Furthermore, replacing the *Paired-stream Embedding Setup* with a single embedding model causes moderate declines, confirming the benefit of combining complementary embeddings to stabilize performance. Finally, replacing the *Optimized Hybrid Semantic Fusion* with simple averaging yields inferior results, underscoring its importance in integrating supervised and unsupervised signals into a unified score.

Method	Mac-P	Mac-R	Mac-F1	Mic-F1
ReDSM5 (Bao et al., 2025)				
SVM†	–	–	28.00	39.00
CNN†	–	–	19.00	25.00
BERT†	–	–	36.00	51.00
LLaMA-3.2-1B†	–	–	49.00	54.00
HOPE (Our)	**72.76** ±1.54	**80.65** ±0.71	**76.15** ±0.50	**77.87** ±0.32
PRIMATE (Gupta et al., 2022)				
BERT‡	–	–	66.43	75.48
MentalBERT‡	–	–	65.37	75.82
SpanEmo‡	–	–	66.25	75.36
LEAR‡	–	–	67.25	75.83
FL-BERT‡	–	–	67.98	75.66
aMLP‡	–	–	64.28	73.21
LR-GCN‡	–	–	66.95	74.97
SpanPHQ‡	–	–	68.84	75.92
PSAT§	–	–	61.60	–
KiNN2¶	–	–	56.00	–
HOPE (Our)	**73.02** ±1.72	**69.81** ±0.57	**70.63** ±0.48	**77.74** ±0.58
DepressionEmo (Rahman et al., 2024)				
SVM*	72.00	41.00	47.00	61.00
Light GBM*	48.00	80.00	58.00	65.00
XGBoost*	63.00	56.00	59.00	66.00
GAN-BERT*	69.00	72.00	70.00	75.00
BERT*	72.00	77.00	74.00	79.00
BART*	70.00	**81.00**	76.00	80.00
Stacked LSTM††	**80.00**	63.00	70.00	–
SelfAug-CL‡‡	76.40	80.31	78.16	–
RoBERTa§§	–	–	–	81.00
HOPE (Our)	78.74 ±0.50	80.75 ±0.59	**79.69** ±0.21	**82.81** ±0.14

Mac: Macro-averaged, Mic: Micro-averaged.
*Best scores in **bold**, second best underlined*, –: Results not reported.
Our results are mean ± standard deviation over 10 runs.
† (Bao et al., 2025), ‡ (Zhang et al., 2023), § (Dalal et al., 2025), ¶ (Dalal et al., 2024), * (Rahman et al., 2024), †† (Younas et al., 2025), ‡‡ (Khan et al., 2024), §§ (Violides et al., 2024).

Table 3: Classification-based evaluation results for our method on three datasets (%).

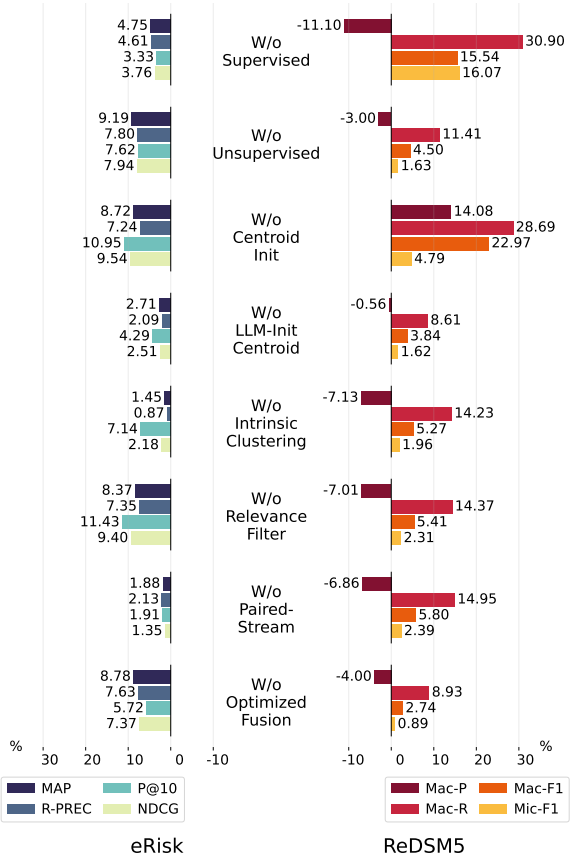

Figure 3: Model Component Contribution Analysis. *W/o: Without, Mac: Macro-averaged, Mic: Micro-averaged.*

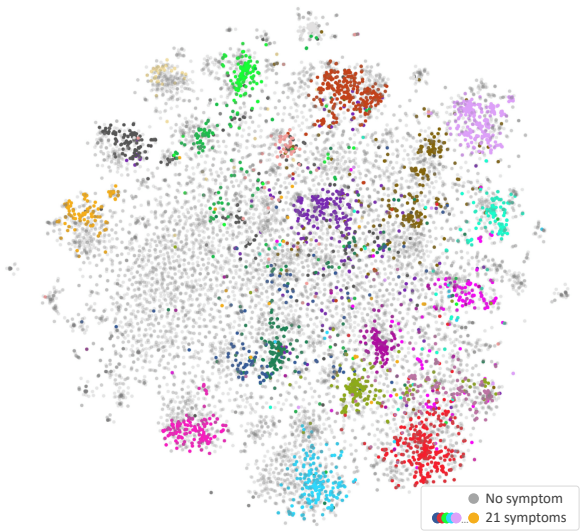

Figure 4: Sentence clustering visualization of eRisk's training data using t-SNE dimension reduction.

5 Discussion

This section discusses the key insights from semantic clustering that motivate the hybrid architecture and relevance filtering. We also present an error analysis highlighting the method's limitations.

5.1 Insights from Semantic Clustering for Hybrid Architecture and Relevance Filter

Figure 4 shows a representative clustering of sentences from the eRisk training set, with 21 depression symptoms and a non-relevant class. We use this figure as a representative case given the dataset's scale and fine-grained taxonomy; similar structures were observed on the other datasets by adapting the descriptor sets for DSM-5, PHQ-9, and emotion labels. The visualization reveals several insights that support our methodological design choices.

Rationale for Hybrid Architecture. The visualization reveals the inherent challenges of multi-label depression symptom classification with several misclassified points and sentences positioned in the boundary regions between clusters. This phenomenon illustrates the difficulty of assigning definitive labels to sentences that may express multiple symptoms simultaneously or exhibit ambiguous symptom associations. Such observations justify our optimized hybrid fusion mechanism, which improves upon purely unsupervised clustering in these complex multi-label contexts.

Rationale for Relevance Filter. The 21 symptom classes exhibit relatively distinctive clustering patterns in the reduced semantic space. Well-formed clusters demonstrate both strong internal cohesion and noticeable separation from other symptoms, supporting our hypothesis that depression symptoms can be effectively modeled through embedding-based clustering.

In contrast, the non-relevant class displays a varied, unpredictable, and scattered distribution throughout the semantic vector space. This scattered nature of non-relevant sentences validates the necessity of our relevance filtering strategy, as these non-relevant sentences lack a coherent semantic structure and would introduce significant noise into the clustering process if not pre-filtered.

5.2 Error Analysis

We conducted a qualitative error analysis to better understand the limitations of the HOPE model. Table 4 presents the most frequent types of prediction errors and their underlying causes, providing insights into the challenges inherent in fine-grained depression symptom detection.

Our analysis identifies four primary categories

Label	Prediction	Cause	Examples
Self-Criticalness	Self-Dislike	Confusion among similar symptoms	*I hate that I feel this way about myself, but I guess it's the circumstance I was raised in.*
Changes in Sleeping Pattern	Tiredness or Fatigue	Missing symptom	*I'm tired and I haven't slept in days*
Self-Dislike	Self-Criticalness	Wrong label annotation	*I'm hard on myself in almost everything that I do.*
Loss of Pleasure	Loss of Interest	Wrong label annotation	*I am just losing interest, that's all.*
Changes in Sleeping Pattern	Tiredness or Fatigue	Missing symptom	*I sleep almost 15 hours a day because I'm physically, mentally and emotionally exhausted.*
Loss of Pleasure, Loss of Interest	Loss of Pleasure	Missing symptom	*I used to enjoy my work etc.. now I no longer do.*
Past Failure, Pessimism	Past Failure	Missing symptom	*I constantly fail at getting what I want in life, no matter how hard I try.*
Self-Dislike, Worthlessness, Past Failure	Self-Dislike, Self-Criticalness, Worthlessness	Wrong label annotation	*I hate myself for being so weak.*

Sentences are extracted from the Reddit-sourced eRisk dataset.

Table 4: Representative mis-classification examples and underlying causes of the proposed model

of model errors. **Confusion among similar symptoms** is the most significant challenge, where the model struggles to distinguish between semantically overlapping symptoms. For instance, the model frequently confuses "Self-Criticalness" with "Self-Dislike" as both symptoms share overlapping linguistic patterns related to negative self-evaluation. The example *"I hate that I feel this way about myself, but I guess it's the circumstance I was raised in"* demonstrates this ambiguity, as it simultaneously expresses self-critical thoughts and self-dislike sentiments.

Missing symptom detection occurs when the model fails to identify all relevant symptoms present in a sentence, typically due to implicit or indirect symptom expressions. The model particularly struggles when symptoms are described through their consequences rather than direct statements, such as *"I sleep almost 15 hours a day because I'm physically, mentally and emotionally exhausted"*, where the change in sleep patterns is justified by fatigue rather than explicit sleep-disturbance language.

Annotation ambiguity highlights cases where ground-truth labels may be inconsistent or where multiple valid interpretations exist. For example, the sentence *"I'm hard on myself in almost everything that I do"* could reasonably be labeled as either "Self-Dislike" or "Self-Criticalness", suggesting inherent ambiguity in human annotation that reflects the complex and overlapping nature of depression symptoms. Such instances underscore the significant challenges of achieving perfect consensus in subjective psychological data labeling.

The analysis also reveals that multi-label sentences pose particular challenges. When sentences express multiple symptoms simultaneously, such as *"I hate myself for being so weak"* (expressing "Self-Dislike", "Worthlessness", and "Past Failure"), the model tends to predict only a subset of the relevant labels. This indicates that while our fusion strategy is effective overall, there remains room for improvement in handling high-density, multi-symptom cases.

6 Conclusion

This work addresses depressive symptom detection on social media with HOPE (Hybrid Optimized Parallel Encoding), a novel framework that simultaneously exploits supervised patterns from labeled training data and unsupervised semantic clustering from unlabeled testing data in a transductive setting. By integrating clinical knowledge through few-shot-enhanced centroid initialization, our unsupervised module learns semantically meaningful representations, enabling automatic class-anchor adjustment and mitigating the temporal domain shift between training and testing data. The optimized hybrid fusion mechanism dynamically integrates signals from both paradigms via a learnable module. Our model achieves robust performance across multiple datasets in both ranking and classification tasks, demonstrating a promising approach for reliable mental health detection systems.

Limitations

Despite the promising results, our approach has several limitations that warrant consideration for future research. First, our model operates primarily at the sentence level and does not account for inter-sentential context, discourse structure, or longitudinal user timelines. Given that many symptom cues are distributed across posts or evolve over time (e.g., worsening mood, shifts in sleep or appetite), incorporating temporal dynamics and user-level context could significantly enhance detection accuracy.

Second, the model occasionally fails to distinguish between symptoms with overlapping linguistic patterns within clinical frameworks (BDI-II, DSM-5, PHQ-9, and emotion labels). This indicates a need for more sophisticated semantic differentiation mechanisms to handle fine-grained clinical nuances.

Third, our current implementation relies on the pre-trained DepRoBERTa tokenizer, which may not fully capture the complexities of social media language such as spelling variations, informal slang, and emojis. In practical applications, these challenges may require domain-aware preprocessing techniques such as slang normalization, emoji interpretation, and multilingual token alignment to ensure more robust detection.

Furthermore, our evaluation focuses on Reddit. Generalization to other platforms, demographics, and cultural settings remains to be established. Future work should incorporate robust normalization, multilingual adaptation, cross-platform and cross-cultural transfer learning to enhance the external validity of the HOPE framework.

Acknowledgments

This work was supported by the Vietnam National Foundation for Science and Technology Development (NAFOSTED) under the project "Research and Development of a Personalized Machine Learning System for Early Diagnosis of Alzheimer's Disease Using Adaptive Multimodal Biomarkers" (Grant No. 102.05-2025.54).

We gratefully acknowledge Kaggle[3] for providing free cloud-based computational resources that supported the implementation and experimental evaluation of this work.

[3] https://www.kaggle.com/

References

Ali Abdulkarem Habib Alrammahi, Farah Abbas Obaid Sari, Zahraa Azhar Muhammad, Mustafa Noaman Kadhim, Dhiah Al-Shammary, and Ayman Ibaida. 2025. Enhancing spam detection with advanced feature extraction and unsupervised clustering. *International Journal of Information Technology*, pages 1–11.

Ling An, Haibo Hu, Mengke Song, Lin Cheng, Shuo Ba, Zhaocong Liu, Zhuohang Yu, Zhenyu Zhang, Yi Liu, and Chichun Zhou. 2025. Unsupervised classification for circulating tumor cells. *IEEE Access*, 13:85669–85681.

Eliseo Bao, Anxo Pérez, and Javier Parapar. 2024. Explainable depression symptom detection in social media. *Health Information Science and Systems*, 12(1):47.

Eliseo Bao, Anxo Pérez, and Javier Parapar. 2025. ReDSM5: A Reddit Dataset for DSM-5 Depression Detection. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 6323–6327.

AP Bascuñana and Isabel Segura Bedmar. 2024. APB-UC3M at eRisk 2024: natural language processing and deep learning for the early detection of mental disorders. *Working Notes of CLEF*, pages 9–12.

AT Beck, CH Ward, M Mendelson, J Mock, and J Erbaugh. 1961. An inventory for measuring depression. *Archives of general psychiatry*, 4(6):561–571.

Ana-Maria Bucur, Adrian Cosma, Paolo Rosso, and Liviu P Dinu. 2023. It's just a matter of time: Detecting depression with time-enriched multimodal transformers. In *European conference on information retrieval*, pages 200–215, Cham. Springer, Springer Nature Switzerland.

Sofia Chaudhary, Jennifer A Hoffmann, Christian D Pulcini, Mark Zamani, Matt Hall, Kristyn N Jeffries, Rachel Myers, Joel Fein, Bonnie T Zima, Peter F Ehrlich, and 1 others. 2024. Youth suicide and preceding mental health diagnosis. *JAMA network open*, 7(7):e2423996–e2423996.

Sumit Dalal, Sarika Jain, and Mayank Dave. 2024. Deep Knowledge-Infusion For Explainable Depression Detection. *arXiv preprint arXiv:2409.02122*.

Sumit Dalal, Deepa Tilwani, Manas Gaur, Sarika Jain, Valerie L Shalin, and Amit P Sheth. 2025. A Cross Attention Approach to Diagnostic Explainability Using Clinical Practice Guidelines for Depression. *IEEE Journal of Biomedical and Health Informatics*, 29(2):1333–1342.

David Guecha, Aaryan Potdar, and Anthony Miyaguchi. 2024. DS@GT eRisk 2024: Sentence Transformers for Social Media Risk Assessment. *Working Notes of CLEF*, pages 825–833.

Lin Lawrence Guo, Stephen R Pfohl, Jason Fries, Alistair EW Johnson, Jose Posada, Catherine Aftandilian, Nigam Shah, and Lillian Sung. 2022. Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine. *Scientific reports*, 12(1):2726.

Shrey Gupta, Anmol Agarwal, Manas Gaur, Kaushik Roy, Vignesh Narayanan, Ponnurangam Kumaraguru, and Amit Sheth. 2022. Learning to Automate Follow-up Question Generation using Process Knowledge for Depression Triage on Reddit Posts. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 137–147, Seattle, USA. Association for Computational Linguistics.

AbdelMoniem Helmy, Radwa Nassar, and Nagy Ramdan. 2024. Depression detection for twitter users using sentiment analysis in English and Arabic tweets. *Artificial intelligence in medicine*, 147:102716.

Pervaiz Iqbal Khan, Andreas Dengel, and Sheraz Ahmed. 2024. Improving Text Representation for Disease Detection from Social Media via Self-augmentation and Contrastive Learning. In *International Conference on Neural Information Processing*, pages 137–151. Springer.

Andriy Kosar, Guy De Pauw, and Walter Daelemans. 2022. Unsupervised text classification with neural word embeddings. *Computational Linguistics in the Netherlands Journal*, 12:165–181.

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

Siryeol Lee, Juncheol Lee, Juntae Park, Jiwoo Park, Dohoon Kim, Joohyun Lee, and Jaehoon Oh. 2024. Deep learning-based natural language processing for detecting medical symptoms and histories in emergency patient triage. *The American Journal of Emergency Medicine*, 77:29–38.

Zhongchen Ma and Songcan Chen. 2021. Expand globally, shrink locally: Discriminant multi-label learning with missing labels. *Pattern Recognition*, 111:107675.

Tu-Phuong Mai, Minh-Ha H Le, Duc-Luong Tran, Duy-Cat Can, and Hoang-Quynh Le. 2025. UET@ eRisk2025: Severity Estimation for Depression Symptoms Searching and Early Risk Detection. *Working Notes of CLEF*, pages 9–12.

Kirill Milintsevich, Gaël Dias, and Kairit Sirts. 2024. Evaluating Lexicon Incorporation for Depression Symptom Estimation. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 322–328, Mexico City, Mexico. Association for Computational Linguistics.

Quang Vinh Nguyen, Dong Thanh Nguyen, Duc Duy Nguyen, Doan Khai Ta, Hai Binh Nguyen, Ji-eun Shin, Seungwon Kim, Hyung-Jeong Yang, and Soo-Hyung Kim. 2025. A Time-Aware Mental State Space for Multimodal Depression Detection on Social Media. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47.

Diogo AP Nunes and Eugénio Ribeiro. 2025. INESC-ID@ eRisk 2025: Exploring Fine-Tuned, Similarity-Based, and Prompt-Based Approaches to Depression Symptom Identification. *Working Notes of CLEF*, pages 1474–1485.

Javier Parapar, Patricia Martín-Rodilla, David E Losada, and Fabio Crestani. 2023. Overview of erisk 2023: Early risk prediction on the internet. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 294–315. Springer.

Javier Parapar, Patricia Martín-Rodilla, David E. Losada, and Fabio Crestani. 2024. Overview of eRisk 2024: Early Risk Prediction on the Internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 73–92, Cham. Springer Nature Switzerland.

Javier Parapar, Anxo Perez, Xi Wang, and Fabio Crestani. 2025a. eRisk 2025: contextual and conversational approaches for depression challenges. In *European Conference on Information Retrieval*, pages 416–424. Springer.

Javier Parapar, Anxo Perez, Xi Wang, and Fabio Crestani. 2025b. Overview of erisk 2025: Early risk prediction on the internet. In *International conference of the cross-language evaluation forum for European languages*, pages 242–265. Springer.

Rafał Poświata and Michał Perełkiewicz. 2022. OPI@ LT-EDI-ACL2022: Detecting signs of depression from social media text using RoBERTa pre-trained language models. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 276–282.

Abu Bakar Siddiqur Rahman, Hoang-Thang Ta, Lotfollah Najjar, Azad Azadmanesh, and Ali Saffet Gönul. 2024. DepressionEmo: A novel dataset for multilabel classification of depression emotions. *Journal of Affective Disorders*, 366:445–458.

Tim Schopf, Daniel Braun, and Florian Matthes. 2022. Evaluating unsupervised text classification: zero-shot and similarity-based approaches. In *Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval*, pages 6–15.

Nguyen Minh Son and DV Thin. 2025. Sonuit eRisk2025: enhanced depression detection on social media via filtering and re-ranking. *Working Notes of CLEF*, pages 9–12.

Matthew Squires, Xiaohui Tao, Soman Elangovan, Raj Gururajan, Xujuan Zhou, U Rajendra Acharya, and Yuefeng Li. 2023. Deep learning and machine learning in psychiatry: a survey of current progress in

depression detection, diagnosis and treatment. *Brain Informatics*, 10(1):10.

Phi Ta, Nha Tran, Hung Nguyen, and Hien D Nguyen. 2025. Detecting signs of depression on social media: A machine learning analysis and evaluation. *Sustainable Futures*, page 100827.

Vankayala Tejaswini, Korra Sathya Babu, and Bibhudatta Sahoo. 2024. Depression detection from social media text analysis using natural language processing techniques and hybrid deep learning model. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(1):1–20.

Poojan Vachharajani. 2025. Transformer ensembles and llm-powered approaches for depression symptom analysis and contextualized early risk detection. *Working Notes of CLEF*, pages 9–12.

Vihang N Vahia. 2013. Diagnostic and statistical manual of mental disorders 5: A quick glance. *Indian journal of psychiatry*, 55(3):220–223.

Marc Violides, Tanatip Timtong, Krystof Bezdek, and Pavlos Andreadis. 2024. Impact of COVID-19 on linguistic expression of depression in online communities. In *Proceedings of the 2024 5th International Symposium on Artificial Intelligence for Medicine Science*, pages 787–793.

Qin Wang, Cees Taal, and Olga Fink. 2021. Integrating expert knowledge with domain adaptation for unsupervised fault diagnosis. *IEEE Transactions on Instrumentation and Measurement*, 71:1–12.

Yuxi Wang, Diana Inkpen, and Prasadith Kirinde Gamaarachchige. 2024. Explainable depression detection using large language models on social media data. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 108–126.

Yutong Wu, David Conlan, Siegfried Perez, and Anthony Nguyen. 2024. Leveraging Unlabeled Clinical Data to Boost Performance of Risk Stratification Models for Suspected Acute Coronary Syndrome. In *AMIA Annual Symposium Proceedings*, volume 2023, page 744.

Aqsa Younas, Shazia Riaz, Saqib Ali, Rafiullah Khan, Mohib Ullah, and Daehan Kwak. 2025. Stacked LSTM Model for Contextual Correlation Detection Among Multiple Emotions. *IEEE Access*, 13:117558–117570.

Tao Yu, Wei Huang, Xin Tang, and Duosi Zheng. 2025. A hybrid unsupervised machine learning model with spectral clustering and semi-supervised support vector machine for credit risk assessment. *PloS one*, 20(1):e0316557.

Tianlin Zhang, Kailai Yang, Hassan Alhuzali, Boyang Liu, and Sophia Ananiadou. 2023. PHQ-aware depressive symptoms identification with similarity contrastive learning on social media. *Information Processing & Management*, 60(5):103417.

Ying Zhang, Xiaocan Jia, Yongli Yang, Na Sun, Shuyan Shi, and Wei Wang. 2024. Change in the global burden of depression from 1990-2019 and its prediction for 2030. *Journal of psychiatric research*, 178:16–22.

Zhiling Zhang, Siyuan Chen, Mengyue Wu, and Kenny Zhu. 2022. Symptom Identification for Interpretable Detection of Multiple Mental Disorders on Social Media. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9970–9985, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hamad Zogan, Imran Razzak, Shoaib Jameel, and Guandong Xu. 2021. Depressionnet: learning multimodalities with user post summarization for depression detection on social media. In *proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 133–142.

A Supplementary Materials Availability Statement

- Source code is available from `https://github.com/candleMind/hope`.
- The eRisk dataset is available upon request to the eRisk organizers at `https://erisk.irlab.org/eRisk2025.html`.
- The ReDSM5 dataset is hosted on Hugging Face `https://huggingface.co/datasets/irlab-udc/redsm5` and requires requesting gated access through the platform.
- The PRIMATE dataset is hosted at `https://github.com/primate-mh/Primate2022` and requires signing the PRIMATE dataset agreement form to obtain access.
- The DepressionEmo dataset is publicly available at `https://github.com/abuBakarSiddiqurRahman/DepressionEmo`.

B Training Environment and Hyperparameters Configurations

This appendix details the computational environments, software stacks, and hyperparameter configurations used for training and evaluating all baseline models.

B.1 Hardware and Software Environment

Experiments were conducted in the free Kaggle Notebook GPU environment. The exact hardware configuration may vary depending on Kaggle resource availability. According to the official Kaggle documentation, free Notebook GPU sessions

provide a single NVIDIA Tesla P100 GPU, while CPU and main memory may vary across sessions.

During our runs in October 2025, the runtime environment reported the following configuration:

- Operating system: Ubuntu 22.04 LTS
- CPU: Intel(R) Xeon(R) CPU @ 2.20GHz
- RAM: 32GB
- GPU: NVIDIA Tesla P100 (16GB VRAM)

The software stack used in our experiments was:

- Python: 3.11.11
- CUDA: 12.6
- PyTorch: 2.6.0+cu124
- HuggingFace Transformers: 4.51.3
- SentenceTransformers: 3.4.1
- PEFT: 0.14.0
- spaCy: 3.8.5, with `en_core_web_sm` for sentence segmentation
- Additional Python packages: `datasets` 3.6.0, `scikit-learn` 1.2.2, `pandas` 2.2.3, `scipy` 1.15.2, `numpy` 1.26.4, `matplotlib` 3.7.2, `plotly` 5.24.1, `beautifulsoup4` 4.13.3, and `pytrec_eval` 0.5

B.2 Model-Specific Hyper-parameters

B.2.1 Symptom Extractor (DepRoBERTa)

The following lists the base model configuration:

- HuggingFace DepRoBERTa's checkpoint path: `rafalposwiata/deproberta-large-depression`.
- Model size: 355M params
- Number of Transformer layers: 24
- Number of Attention heads: 16
- Hidden size: 1024
- Intermediate Size: 4096

The parameter-efficient fine-tuning (LoRA) configuration is summarized below:

- Rank (r): 32
- Scaling factor (α): 64
- Dropout: 0.1
- Target modules: query, value

Input preprocessing and tokenization are configured as follows:

- Tokenizer: `rafalposwiata/deproberta-large-depression`
- Max length: 256 tokens
- Truncation: True
- Padding: max length

The optimization parameters are listed below:

- Optimizer: AdamW
- Learning rate: 1e-5
- LR scheduler: `CosineAnnealingLR`
- Warmup ratio: 0.1
- Weight decay: 0.01

The training setup is summarized as follows:

- Batch size: 32
- Epochs: 20
- Loss function: `BCEWithLogitsLoss`
- Problem type: Multi-label classification

B.2.2 Unsupervised Intrinsic Semantic Clustering

The following describes the configuration for the **Pair-stream Semantic Clustering** module:

- Embedding models:
  - `nomic-ai/modernbert-embed-base`
  - `nomic-ai/nomic-embed-text-v1.5`
- Clustering algorithm: K-Means
- Initialization: centroids initialized with symptom description embeddings
- Few-shot seeding examples: 5
- Max iterations: 300
- Convergence tolerance: 1e-4

The configuration for the **Pair-stream Aggregation** module is summarized below:

- Fusion strategy: average ensemble
- Aggregation unit: label-score vector
- Output representation: mean vector of two label-score vectors

B.2.3 Optimized Hybrid Semantic Fusion

- Fusion strategy: vector concatenation followed by a multi-layer perceptron (MLP)
- Input dimension: $2 \times L$ (where L is the number of labels)
- Hidden layers: `[64]`
- Activation function: `ReLU`
- Dropout: 0.1
- Loss function: `BCEWithLogitsLoss`
- Optimizer: `AdamW`
- Learning rate: 2e-3
- LR scheduler: `CosineAnnealingLR`
- Training epochs: 100
- Batch size: 32

B.3 Training Time

On average, training modules in our model take:

- **Symptom Extractor (DepRoBERTa)**: 4-10.5 hours depending on the dataset size (eRisk, ReDSM5, PRIMATE, DepressionEmo).
- **Unsupervised Intrinsic Semantic Clustering**: approximately 10-30 minutes for computing sentence embeddings and running clustering (depending on corpus size and number of clusters).
- **Optimized Hybrid Semantic Fusion**: approximately under 3-5 minutes for training the final MLP on concatenated representations.

C Symptoms Description

This section provides an overview of the depressive symptoms included in our study across four datasets: eRisk 2025 Task 1 (BDI-II symptoms), ReDSM5 (DSM-5 symptoms), PRIMATE (PHQ-9 symptoms), and DepressionEmo. These serve as symptom-descriptive inputs before the clustering stage, providing semantic initialization for the Symptom-Initialized K-means module. The symptom lists for each dataset are presented below.

1. **BDI-II Symptoms**
   The 21 BDI-II symptoms included in this study are: Sadness, Pessimism, Past Failure, Loss of Pleasure, Guilty Feelings, Punishment Feelings, Self-Dislike, Self-Criticalness, Suicidal Thoughts or Wishes, Crying, Agitation, Loss of Interest, Indecisiveness, Worthlessness, Loss of Energy, Changes in Sleeping Pattern, Irritability, Changes in Appetite, Concentration Difficulty, Tiredness or Fatigue, Loss of Interest in Sex.

2. **DSM-5 Symptoms**
   The nine DSM-5 symptoms included in this study are: Anhedonia, Appetite Change, Cognitive Issue, Depressed Mood, Fatigue, Psychomotor Alteration, Sleep Issues, Suicidal Thoughts, Worthlessness.

3. **PHQ-9 Symptoms**
   The nine PHQ-9 symptoms included in this study are: Hyper/Lower Activity, Concentration Problem, Eating Disorder, Feeling Down, Lack of Energy, Lack of Interest, Low Self-Esteem, Self-Harm, Sleeping Disorder.

4. **DepressionEmo's Symptoms**
   The eight DepressionEmo's symptoms included in this study are: Sadness, Suicide intent, Worthlessness, Anger, Cognitive Dysfunction, Empty, Hopelessness, Loneliness.

An example of one symptom from the BDI-II, which includes seven statements describing different levels and manifestations of that symptom:

Changes in Sleeping Pattern:

1. I have not experienced any change in my sleeping.
2. I sleep somewhat more than usual.
3. I sleep somewhat less than usual.
4. I sleep a lot more than usual.
5. I sleep a lot less than usual.
6. I sleep most of the day.
7. I wake up 1–2 hours early and can't get back to sleep.

The complete lists and detailed descriptions for each symptom across all datasets are publicly available at:

- BDI-II Questionnaire (Beck et al., 1961): https://naviauxlab.ucsd.edu/wp-content/uploads/2020/09/BDI21.pdf
- DSM-5 Symptoms Definitions (Vahia, 2013): https://floridabhcenter.org/wp-content/uploads/2021/03/MDD_Adult-Guidelines-2019-2020.pdf
- PHQ-9 Questionnaire (Kroenke et al., 2001): https://www.apa.org/depression-guideline/patient-health-questionnaire.pdf
- DepressionEmo's Symptoms Definitions (Rahman et al., 2024): https://arxiv.org/pdf/2401.04655

Note that BDI-II and PHQ-9 are standardized self-report questionnaires, whereas DSM-5 and DepressionEmo provide only conceptual or dataset-specific definitions of depressive symptoms. In particular, DepressionEmo represents the symptom definitions manually established by the authors in this study.