scientific reports

Check for updates

OPEN

Multi-label mental health classification in social media posts with multi-perspective prompt ensemble and auxiliary self-supervision

Cheng-Ying Hsieh[1], Qing-Yuan Ye[2], Feng-Chi Liu[3], Xin Wang[4], Cheng-Hsiung Lee[2] & Ching-Sheng Lin[2]✉

Anxiety and depression have become major global health concerns. With the rapid rise of social media, people increasingly share emotions and personal struggles through posts, which often convey multiple mental states simultaneously. To address this multi-label classification challenge in mental health texts, this study proposes a multi-task framework with two main modules, a multi-perspective prompt design module and a perturbation-based self-supervised learning module, based on a pre-trained language model backbone. Prompts from sociological, psychological, and educational perspectives are used to enhance semantic understanding. To improve model robustness, we formulate self-supervised auxiliary tasks where the model predicts whether a sentence has undergone insertion, swap, or deletion. Experiments on the MultiWD dataset, covering six wellness dimensions, show that our method outperforms all baselines. Furthermore, ablation studies explore the impact of different training configurations and confirm the critical contributions of both proposed modules.

Keywords Multi-label classification, Mental health, Multi-perspective prompt, Perturbation-based self-supervised learning, Pre-trained language model

Mental disorders globally affect a substantial portion of the population, with lifetime prevalence estimated at up to 50%. Recent research over the past decade highlights a rising trend in cases across all age groups, including children, adolescents, and adults[1]. Social media, now deeply embedded in everyday life, has become a public outlet for expression. Individuals, especially those facing emotional or psychological struggles, often turn to these platforms to share their thoughts[2]. These user-generated texts offer rich insights for understanding mental health conditions. Nevertheless, due to the exponential increase in social media content, conducting mental health evaluations manually is no longer feasible. As a result, researchers have turned to natural language processing (NLP) techniques for scalable and automatic analysis[3].

Previous studies on mental health detection have primarily focused on traditional machine learning approaches[4]. However, these methods face several limitations. First, traditional models often struggle to capture the nuanced semantics and context-dependent expressions characteristic of mental health-related texts[5]. In addition, manual feature extraction is time-consuming, requires domain expertise, and often brings limited performance gains[6]. Building on the transformative advancements deep learning has brought to NLP, researchers have increasingly adopted these approaches to confront this challenge. Long Short-Term Memory (LSTM) networks[7,8] and Convolutional Neural Networks (CNN)[9,10] have been widely applied to improve performance in text classification tasks. These approaches are effective at capturing sequential features and localized textual patterns. However, models with relatively simple architectures may lack the ability to capture deeper semantic structures. This limitation becomes especially problematic in mental health classification, where emotional cues are often subtle, indirect, and heavily dependent on context.

[1]Department of Pharmacology, School of Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan. [2]Master Program of Digital Innovation, Tunghai University, Taichung City 40704, Taiwan. [3]Department of Statistics, Feng Chia University, Taichung 40724, Taiwan. [4]College of Integrated Health Sciences and the AI Plus Institute, State University of New York, Albany, NY 12222, USA. ✉email: cslin612@thu.edu.tw

Transformer-based architectures have revolutionized NLP by capturing complex dependencies and contextual nuances across a wide range of tasks. Among them, non-autoregressive methods such as BERT-based models have demonstrated good performance in mental health text classification[11]. However, despite their success, these models tend to underperform when faced with tasks or domains that differ from those seen during pre-training[12]. Generative AI models such as ChatGPT and LLaMA have recently achieved remarkable success across various NLP tasks and demonstrated powerful capabilities in text generation, reasoning, and interaction. These models have also been widely explored in the domain of mental illness classification, where their in-context learning ability combined with prompt engineering and expert-crafted few-shot examples has shown promising results. However, their performance still falls short when compared to non-autoregressive and domain-specific models such as MentalRoBERTa[13,14].

Building on prior observations, conventional pre-trained language models (PLMs), such as BERT, often fail to generalize well to novel mental health-related tasks. Generative AI models offer strong in-context learning but face challenges such as instability and hallucinations, especially in sensitive domains. To address these issues, we propose a multi-perspective prompt architecture with self-supervised learning that models psychological cues from multiple semantic views. Moreover, the self-supervised objective enables the model to learn rich and transferable representations without relying on costly annotations, thereby enhancing generalization and stability across diverse mental health scenarios.

The primary contributions of our research are summarized below:

- We propose an ensemble framework for multi-label mental health classification in social media posts, which integrates multiple persona-specific prompts into a non-autoregressive pre-trained language model. To enhance robustness and representation learning, we introduce a self-supervised auxiliary task that predicts semantic perturbations of the input text.
- The proposed approach achieves competitive results on the public benchmark MultiWD for mental health classification. Ablation experiments also validate the contribution of each module to the overall performance. Moreover, although our study utilizes the MultiWD dataset which is derived from Reddit posts, the proposed model is expected to be seamlessly extendable to different social media platforms due to the similarly unstructured and user-generated nature of textual content across these communities.

The remainder of this paper is organized as follows. Section 2 provides a review and discussion of related technologies and relevant works. In Sect. 3, we present our proposed model architecture. Section 4 outlines our experimental studies and their results. Finally, Sect. 5 concludes the paper and discusses potential future research directions.

Related works

The growing concern over mental health issues has driven a surge of research leveraging textual data for automated detection and analysis. Prior studies range from traditional machine learning methods to advanced deep learning architectures and language models. In the following sections, we provide a structured review of these approaches.

Traditional machine learning methods have played a foundational role in early research efforts to detect mental health problems in text. Support Vector Machines (SVM) emerges as a preferred classification method due to their strong performance on high-dimensional data. SVM with a radial basis function kernel has been applied to detect depression on Twitter by analyzing language use and behavioral signals, showing that depressed users often display distinct linguistic patterns and lowered social activity[15]. A decision tree framework with profile-based and sentiment-aware design is proposed for depression detection in social media, where classifiers are customized based on user demographics and enhanced with dual polarity-aware bag-of-words representations[16]. Going beyond the limitations of traditional latent Dirichlet al.location (LDA), a supervised nested LDA model (SNLDA) identifies depression-related language on Twitter by jointly modeling topic distributions and label information. It leverages hierarchical topic structures to capture subtle semantic patterns and leads to improved classification accuracy and better interpretability[17].

Prior to the widespread adoption of transformer-based and language models, deep learning methods such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory networks (LSTMs) are widely used in mental health classification tasks on social media data. These models automatically learn hierarchical or sequential features from raw text without relying on manually crafted features. A CNN-based deep learning model classifies users' mental states by analyzing their posting history on Reddit[10]. By training on data from mental health-related communities, the model effectively detects six specific mental disorders, offering a potential supplementary tool for large-scale mental health monitoring via social media. XA-BiLSTM integrates XGBoost for data balancing and an attention-enhanced BiLSTM for classification, and it achieves superior performance compared to prior models on the RSDD dataset[18]. A stacked CNN followed by a two-layer LSTM architecture predicts suicidal ideation on social media by capturing both local textual features and long-term dependencies[19]. Experiments demonstrate that the model achieves high classification accuracy and outperforms previous CNN-LSTM baselines.

Transformer-based architectures have demonstrated great success in numerous domains, particularly in language modeling, where they can be broadly divided into non-autoregressive and autoregressive models. The BERT-ATT is an attention-based classification framework that leverages BERT-derived contextual embeddings to detect psychiatric disorders from Reddit posts[20]. The model demonstrates superior performance over traditional baselines across both user-level and post-level classification tasks for eight mental health conditions. Another study enhances the generalizability of BERT-based depression detection models by grounding predictions in clinically validated symptoms from the PHQ9 questionnaire. Experimental results across multiple social media

datasets show that this clinically informed approach improves out-of-distribution performance and offers greater model interpretability[21]. The success of generative large language models (LLMs) has recently extended to mental health analysis, motivating their use in tasks like symptom detection and supportive dialogue. ChatGPT (gpt-3.5-turbo) is evaluated in zero-shot settings on three mental health classification tasks—stress, depression, and suicidality detection—using annotated social media datasets[22]. The model achieves promising F1 scores and significantly outperforms majority-class baselines, highlighting its potential for mental health applications. MentaLLaMA introduces an open-source instruction-tuned LLM designed for interpretable mental health analysis on social media, trained on a newly constructed IMHI dataset containing 105 K multi-task and multi-source samples[12]. Experimental results show that MentaLLaMA achieves performance comparable to state-of-the-art discriminative models while generating human-level explanations. To address demographic bias in LLM-based mental health analysis, a comprehensive evaluation reveals that GPT-4 achieves the best balance of performance and fairness[14]. This concern with mitigating bias and enriching representations across heterogeneous data domains is also reflected in recent advances in semi-supervised domain adaptation (SSDA), such as the EnSR framework[23].

Methodology

In this section, we begin by formulating the task of identifying mental health conditions in social media posts as an ensemble-based multi-label problem. Next, our multi-task framework is introduced, leveraging a PLM with multi-perspective prompts and perturbation-based self-supervised learning.

Problem formulation

We formulate the problem as a multi-label classification task, where the goal is to predict a set of labels $\mathrm{y} = [\mathrm{y}_1, \mathrm{y}_2, \ldots \mathrm{y}_\mathrm{C}]$ for a given input social media post x. Each label $\mathrm{y}_\mathrm{c} \in \{0, 1\}$ indicates the presence or absence of the $\mathrm{c}-\mathrm{th}$ category, and C is the total number of possible labels. We leverage multiple prompts $\{\mathrm{p}_1, \mathrm{p}_2, \ldots \mathrm{p}_\mathrm{M}\}$ to capture diverse perspectives of x and incorporate a perturbed version $\widetilde{\mathrm{x}}$ to to enable auxiliary self-supervised learning. This can be formulated as:

$$\widehat{\mathrm{y}^{(\mathrm{m})}}, \widehat{\mathrm{z}^{(\mathrm{m})}} = \mathrm{f}_\theta \left(\mathrm{p}_\mathrm{m}(\mathrm{x}), \widetilde{\mathrm{x}} \right), \text{ where } \mathrm{m} = 1,2 \ldots \mathrm{M} \tag{1}$$

where $\mathrm{f}_\theta(\cdot)$ represents a PL backbone parameterized by θ, $\widehat{\mathrm{y}^{(\mathrm{m})}}$ is the multi-label prediction and $\widehat{\mathrm{z}^{(\mathrm{m})}}$ denotes the auxiliary prediction. The final prediction $\widehat{\mathrm{y}}$ is obtained by aggregating the predictions from all $\widehat{\mathrm{y}^{(\mathrm{m})}}$ via voting scheme.

Proposed approach

We propose a multi-task learning framework based on a PLM to address the problem of multi-label classification in social media posts, as illustrated in Fig. 1. The framework consists of two main modules: the first is the primary multi-label (ML) classification module, which leverages prompt engineering by incorporating various persona prompts from different professional perspectives to capture semantic diversity and enhance classification performance. The second is an auxiliary self-supervised (SS) learning module, which applies perturbations to the input post to generate corrupted text variants, and trains the model to identify the type of perturbation, thereby improving its robustness and generalization to semantic variations.

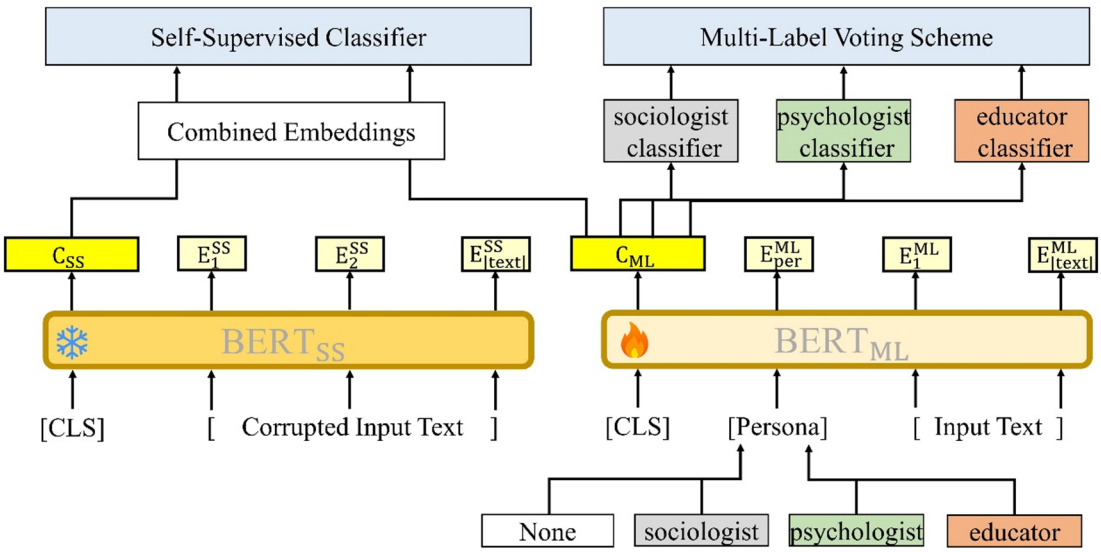

Fig. 1. The system architecture.

Our primary task is multi-label classification of mental health-related dimensions in social media posts, as illustrated in the right part of Fig. 1. To enhance the model's ability to capture diverse user intentions and expressions, we introduce persona-specific prompts inspired by different expert perspectives. Incorporating these prompts provides two key benefits: (1) it encourages the model to attend to different semantic dimensions of the same post, and (2) it offers a form of implicit multi-view learning, enabling the model to generalize better by interpreting inputs from complementary angles. Given an input post x and its corresponding ground-truth multi-label vector $y \in \{0, 1\}^C$ where C is the total number of possible labels, we first construct a set of persona-specific prompts $\{p_1, p_2, p_3\}$, representing three distinct expert perspectives: sociologist, psychologist, and educator (i.e., $M = 3$). Each prompt p_m is concatenated with the input x and passed through a PLM (e.g., a trainable BERT-based backbone denoted as $\mathrm{BERT}_{\mathrm{ML}}$), from which we extract the corresponding [CLS] embedding $C_{\mathrm{ML}}^{(m)}$. These embeddings are then passed through a shared linear classifier with the weight matrix W_{ML} and the bias term b_{ML}, followed by a sigmoid activation, to produce the multi-label prediction for each persona-specific prompt:

$$\widehat{y^{(m)}} = \sigma\left(W_{\mathrm{ML}}^{T} C_{\mathrm{ML}}^{(m)} + b_{\mathrm{ML}}\right),\ m \in \{1,\ 2,\ 3\} \tag{2}$$

To obtain the final prediction $\widehat{y_{\mathrm{ML}}}$, we apply average ensembling across all three prompt-specific outputs:

$$\widehat{y_{\mathrm{ML}}} = \frac{1}{3}\sum_{m=1}^{3} \widehat{y^{(m)}} \tag{3}$$

The model is trained by minimizing the binary cross-entropy (BCE) loss between the ensembled prediction and the ground truth:

$$L_{\mathrm{ML}} = \mathrm{BCE}(\widehat{y_{\mathrm{ML}}}, y) = -\sum_{i=1}^{C} y_{(i)} \log \widehat{y_{\mathrm{ML}(i)}} + (1 - y_{(i)}) \log(1 - \widehat{y_{\mathrm{ML}(i)}}) \tag{4}$$

This training strategy encourages the model to jointly optimize the output representations across all persona prompts, yielding a more robust and comprehensive prediction. Notably, this ensemble strategy is consistently applied during both training and inference phases.

The full set of persona prompts used in this work is summarized in Table 1. Sociological prompts focus on social roles, interactions, structures, and support systems, capturing social-context cues such as isolation or social support. Psychological prompts emphasize emotional states, cognitive patterns, and behavioral traits, allowing detection of affective and cognitive indicators like anxiety, depression, or overthinking. Educational prompts highlight learning environments, knowledge acquisition, and skill development, reflecting self-regulation, coping strategies, and personal growth. By combining these perspectives through prompt ensemble, the model captures complementary semantic features across social, psychological, and educational dimensions, enhancing multi-label mental health classification in social media texts.

While persona-specific prompts provide valuable domain-informed perspectives for improving multi-label classification, they may not fully capture semantic variations or linguistic noise inherent in social media texts. These challenges can degrade the generalizability of the model, especially under limited labeled data. To enhance the model's robustness to semantic perturbations and improve its representation learning, we introduce an auxiliary self-supervised learning module, illustrated in the left half of Fig. 1. By learning to identify these controlled perturbations, the model is encouraged to capture deeper semantic dependencies and contextual relationships, rather than relying solely on surface-level word patterns. This robustness is particularly beneficial for social media texts, which often contain spelling errors, non-standard expressions, and word-order variations. Specifically, we feed the original input text x into the same trainable $\mathrm{BERT}_{\mathrm{ML}}$ used for the main task, but without incorporating any persona prompt. We refer to this input configuration as $m = 0$ and denote the resulting embeddings as $C_{\mathrm{ML}}^{(0)}$. In parallel, we apply one of three perturbations to x, namely insertion, token swap, or deletion, resulting in a corrupted input $\tilde{x}$. This corrupted text is then passed through a frozen PLM (e.g., a BERT-based backbone denoted as $\mathrm{BERT}_{\mathrm{SS}}$) to obtain another [CLS] embedding $C_{\mathrm{SS}}^{(0)}$. We concatenate the two embeddings to form a combined representation:

sociologist	As a sociologist, analyze the following text. Consider social roles (e.g., family member, student, employee), social interactions (e.g., communication, cooperation, conflict), social structures (e.g., institutions, norms), and support systems (e.g., family, peers, community). Then determine which of the following life dimensions are most clearly represented in the text (multiple labels allowed): [Spiritual, Physical, Intellectual, Social, Vocational, Emotional]
psychologist	As a psychologist, analyze the following text. Consider emotional states (e.g., sadness, anxiety, happiness), cognitive patterns (e.g., overthinking, learning, reflection), behavioral traits (e.g., social withdrawal, motivation), and mental processes (e.g., coping, self-regulation). Then determine which of the following life dimensions are most clearly represented in the text (multiple labels allowed): [Spiritual, Physical, Intellectual, Social, Vocational, Emotional]
educator	As an educator, analyze the following text. Consider learning environments (e.g., school, informal settings), knowledge acquisition (e.g., reading, studying), skill development (e.g., communication, critical thinking), and educational goals (e.g., self-improvement, life-long learning). Then determine which of the following life dimensions are most clearly represented in the text (multiple labels allowed): [Spiritual, Physical, Intellectual, Social, Vocational, Emotional]

Table 1. The prompts for different personas.

$$h_{SS} = [C_{ML}^{(0)}; C_{SS}^{(0)}] \tag{5}$$

This representation is then passed through a linear classifier to predict the perturbation type on x,

$$\hat{z} = \mathrm{softmax}(W_{SS}^{T} h_{SS} + b_{SS}) \tag{6}$$

where W_{SS} and b_{SS} are learnable parameters. The self-supervised loss is then defined using the cross-entropy between $\hat{z}$ and the true perturbation label z:

$$L_{SS} = CE(\hat{z}, z) = \sum_{k=1}^{3} z_{(k)} \log \hat{z}_{(k)} \tag{7}$$

Our final training objective combines the losses from the main multi-label classification task and the auxiliary self-supervised perturbation prediction task. The overall training loss is defined as:

$$L_{total} = L_{ML} + \alpha L_{SS} \tag{8}$$

where α is a hyperparameter that balances the contribution of the auxiliary task and we use θ to denote all trainable parameters in optimizing L_{total}. Note that in Fig. 1, E_j^i refers to the embeddings of all tokens except the [CLS] token, which are not used in our method.

Algorithm 1 presents the pseudo code of the training process for our multi-task learning model. The inputs to the algorithm include the training dataset, a trainable BERT model, and a frozen BERT model. The training process outputs the complete set of learned parameters θ for the entire model. Each training iteration consists of three components: the main multi-label classification task (lines 4–9), the auxiliary self-supervised learning task (lines 11–15), and the optimization step (line 17).

Experiments

In this section, we present the dataset, evaluation metrics, experimental results, and ablation studies.

Dataset and evaluation metric

In this study, we adopt the MultiWD dataset[24], which is annotated across six dimensions of well-being. The dataset consists of 2,624 Reddit posts in the training set and 657 posts in the test set, all formulated as a multi-label classification task. Table 2 summarizes the label distribution.

To evaluate the performance of our model, we adopt three standard metrics: Precision, Recall, and F1-score. Precision is used to measure the proportion of correctly predicted labels among all labels assigned by the model, reflecting its exactness. Recall is applied to quantify the proportion of correctly predicted labels among all ground-truth labels, indicating the model's ability to retrieve relevant labels. The F1-score is the harmonic mean of Precision and Recall, providing a balanced measure of the model's accuracy.

Input: Training data {X,Y}, a trainable PLM $BERT_{ML}$, a frozen PLM $BERT_{SS}$
Output: The trained model with the complete set of learned parameters θ.

1: **for** each training iteration:
2: Obtain the batch data {x, y}
3: # Main multi-label classification task
4: **for** m = 1 to 3:
5: Apply the concatenated input [p_m; x] to $BERT_{ML}$ to obtain embeddings $C_{ML}^{(m)}$
6: Compute the prediction based on Eq. (2)
7: **end for**
8: Calculate the average prediction in Eq. (3)
9: Derive the multi-label classification loss L_{ML} in Eq. (4)
10: # Auxiliary self-supervised learning task
11: $\tilde{x}$ = ({Insert, Swap, Delete})(x)
12: Input $\tilde{x}$ to $BERT_{SS}$ to extract embeddings $C_{SS}^{(0)}$
13: Input x to $BERT_{ML}$ to extract embeddings $C_{ML}^{(0)}$
14: Use combined representation, h_{SS}, to generate the prediction in Eq. (6)
15: Derive the self-supervised classification loss L_{SS} in Eq. (7)
16: # Optimization
17: Update the parameters θ through minimizing L_{total} in Eq. (8)
18: **end for**

Algorithm 1. Multi-task learning model

Label	Training data	Testing data
Spiritual	164	37
Physical	725	198
Intellectual	514	137
Social	1,711	419
Vocational	453	97
Emotional	1,336	326

Table 2. Overview of the multiwd dataset.

Category	Model	Precision	Recall	F1-score
General PLMs	BERT	73.74	81.38	76.69
	ALBERT	74.11	75.12	74.26
	DistilBERT	72.95	78.67	75.43
Domain-specific PLMs	ClinicalBERT	70.86	77.10	73.41
	MentalBERT	72.88	80.48	76.19
	PsychBERT	71.87	76.69	73.92
LLMs	GPT-3	**75.13**	76.94	75.94
	GPT-3.5-turbo	69.53	70.53	69.93
	LLAMA2	34.12	26.67	28.94
Persona-informed + Self-supervised	Our model	69.59	**82.29**	**79.18**

Table 3. Experimental results (in %) of different models on the multiwd dataset.

Experimental results and analysis

Our experiments include comparisons with three categories of training-based models. The first category consists of general-purpose PLMs, including BERT[25], ALBERT[26], and DistilBERT[27]. The second category includes domain-specific PLMs tailored for mental health or clinical texts, such as PsychBERT[28], MentalBERT[29], and ClinicalBERT[30]. The third category comprises LLMs, including GPT-3, GPT-3.5-turbo, and LLaMA2.

Table 3 presents the comparison results, where the baseline performances are taken directly from the prior study[24]. We summarize the following key observations. First, our model achieves the highest Recall and F1-score among all baseline categories. Although its precision is comparatively lower, the strong F1-score indicates a favorable trade-off in early detection or screening scenarios where broad coverage is essential. However, we recognize that false positives could still impose unnecessary psychological or resource burdens. Future work will explore adaptive thresholding and confidence calibration techniques to balance recall with precision more effectively. Second, when comparing across model categories, we observe that both general PLMs (74.26% to 76.69%) and domain-specific PLMs (73.41% to 76.19%) consistently achieve F1-scores in the mid-70s range. LLAMA2's poor F1-score (28.94%), compared to GPT-3 and GPT-3.5-turbo, highlights how LLMs' architecture, instruction alignment, and pre-training data crucially affect downstream performance. Third, excluding the notably poor performance of LLAMA2, all models exhibit fairly consistent results, with Precision scores ranging between 69 and 75, and Recall scores between 70 and 82. This indicates that the models maintain a good balance between correctly identifying relevant labels (Recall) and avoiding false positives (Precision). The relatively higher Recall values suggest that these models are particularly effective at capturing relevant mental health indicators, which is critical in mental health classification tasks where missing a true label can be more detrimental than a false alarm. Despite the existence of domain-specific PLMs, the linguistic mismatch between their formally structured pre-training corpora and the informal, emotionally expressive language of social media highlights the advantage of general-purpose models. Models like BERT, pre-trained on diverse and extensive text distributions, exhibit superior robustness and generalization when fine-tuned on noisy, real-world mental health datasets.

To provide a detailed understanding of the model's performance across six different well-being labels, Table 4 presents the per-label classification statistics on the test set, including true positives (TP), false negatives (FN), false positives (FP), true negatives (TN), Precision, Recall and F1-score. Notably, the Social label has the largest amount of training data with 1,711 instances (from Table 2). It also achieves the highest true positive count of 384 and a relatively low false negative count of 35, indicating strong model performance likely supported by ample training data. On the other hand, the Spiritual label has the fewest training samples, with only 164 instances (from Table 2). It yields the lowest true positive count (10) and relatively high false negatives (27), suggesting weaker predictive performance. This contrast indicates that the amount of training data significantly influences the model's ability to accurately classify each well-being dimension.

To better understand the contribution of each component in our approach, we conduct ablation studies by systematically removing the multi-perspective prompting and the self-supervised perturbation mechanism. The results, summarized in Table 5, reveal several key insights. First, the ablation study demonstrates that the combination of multi-perspective prompting and self-supervised perturbations yields the best overall

Label	TP	FN	FP	TN	Precision	Recall	F1-score
Spiritual	10	27	15	605	40.00	27.03	32.26
Physical	148	50	72	387	67.27	74.75	70.81
Intellectual	105	32	56	464	65.22	76.64	70.47
Social	384	35	81	157	82.58	91.65	86.88
Vocational	80	17	43	517	65.04	82.47	72.73
Emotional	272	54	175	156	60.85	83.44	70.38

Table 4. Per-label classification statistics on the test set (TP, FN, FP, TN in counts; Precision, Recall, F1-score in %).

Model	Precision	Recall	F1-score
Our model	69.59	**82.29**	**79.18**
w/o multi-perspective prompt	**74.78**	73.06	73.24
w/o self-supervised	71.60	78.06	74.47

Table 5. Ablation studies results (in %).

performance, with an F1-score of 79.18%, outperforming the settings without either component. Second, comparing the two ablated variants, removing the multi-perspective prompt leads to a larger drop in F1-score. This suggests that the multi-perspective prompt contributes more significantly to the model's overall performance. Third, removing the multi-perspective prompting leads to a noticeable drop in recall (from 82.29% to 73.06%), although precision improves. This indicates that while the model becomes more conservative in making predictions, it may miss relevant instances without diverse persona cues.

Conclusion and future work

In this study, we investigate the task of multi-label mental health classification by leveraging BERT-based language modeling, augmented with multi-persona perspectives and self-supervised perturbations. Our approach is evaluated on the MultiWD dataset, which encompasses six key well-being dimensions. The results demonstrate the effectiveness of incorporating persona diversity and controlled textual variations for improving predictive performance. Additionally, detailed evaluation and ablation studies highlight the contribution of each component in our framework.

Although our method yields encouraging results on the MultiWD dataset, there are several promising directions to extend the current framework. First, while the present approach employs static persona representations, future work could explore dynamic or context-aware persona routing mechanisms, enabling the model to adaptively select or weight personas based on the linguistic and psychological cues within each input. Second, the current self-supervised perturbation strategy relies on token-level modifications. Future research could investigate semantically controlled perturbations, such as those guided by affective lexicons, topic shifts, or syntactic structures, to ensure more meaningful and robust variations. Finally, we aim to explore the applicability of our proposed framework across a wider range of psychological or clinical NLP tasks, thereby assessing its potential to generalize beyond the current setting.

Data availability

(https:/github.com/drmuskangarg/MultiWD) (accessed on 31 January 2025).

Received: 5 August 2025; Accepted: 27 November 2025
Published online: 05 December 2025

References

1. Viana, M. C. et al. Barriers to 12-month treatment of common anxiety, mood, and substance use disorders in the world mental health (WMH) surveys. *Int. J. Mental Health Syst.* **19** (1), 1–18 (2025).
2. Benrouba, F. & Boudour, R. Emotional sentiment analysis of social media content for mental health safety. *Social Netw. Anal. Min.* **13** (1), 17 (2023).
3. Zhai, W. et al. Chinese mentalbert: Domain-adaptive pre-training on social media for Chinese mental health text analysis. *ArXiv Preprint arXiv* 240209151. (2024).
4. Kerasiotis, M., Ilias, L. & Askounis, D. Depression detection in social media posts using transformer-based models and auxiliary features. *Social Netw. Anal. Min.* **14** (1), 196 (2024).
5. Leiva, V. & Freire, A. Towards suicide prevention: early detection of depression on social media. In Internet Science: 4th International Conference, INSCI 2017, Thessaloniki, Greece, November 22–24, 2017, Proceedings 4 (pp. 428–436). Springer International Publishing. (2017).
6. Bokolo, B. G. & Liu, Q. Advanced comparative analysis of machine learning and transformer models for depression and suicide detection in social media texts. *Electronics* **13** (20), 3980 (2024).
7. Shah, F. M. et al. Early depression detection from social network using deep learning techniques. In 2020 IEEE region 10 symposium (TENSYMP) (pp. 823–826). IEEE. (2020).

8. Wu, M. Y., Shen, C. Y., Wang, E. T. & Chen, A. L. A deep architecture for depression detection using posting, behavior, and living environment data. *J. Intell. Inform. Syst.* **54** (2), 225–244 (2020).
9. Yao, H. et al. Detection of suicidality among opioid users on reddit: machine learning–based approach. *J. Med. Internet. Res.* **22**(11), e15293 (2020).
10. Kim, J., Lee, J., Park, E. & Han, J. A deep learning model for detecting mental illness from user content on social media. *Sci. Rep.* **10** (1), 11846 (2020).
11. Cho, H. N. et al. *Task-Specific Transformer-Based Language Models in Health Care: Scoping Review* Vol. 12 (JMIR Medical Informatics, 2024). e49724.
12. Yang, K. et al. MentaLLaMA: interpretable mental health analysis on social media with large language models. In Proceedings of the ACM Web Conference 2024 (pp. 4489–4500). (2024).
13. Yang, K. et al. Towards interpretable mental health analysis with large Language models. *ArXiv Preprint arXiv* 230403347. (2023).
14. Wang, Y. et al. Unveiling and mitigating bias in mental health analysis with large Language models. *ArXiv Preprint arXiv* 240612033. (2024).
15. De Choudhury, M., Gamon, M., Counts, S. & Horvitz, E. Predicting depression via social media. In Proceedings of the international AAAI conference on web and social media (Vol. 7, No. 1, pp. 128–137). (2013).
16. de Jesús Titla-Tlatelpa, J., Ortega-Mendoza, R. M., Montes-y-Gómez, M. & Villaseñor-Pineda, L. A profile-based sentiment-aware approach for depression detection in social media. *EPJ Data Sci.* **10** (1), 54 (2021).
17. Resnik, P. et al. Beyond LDA: exploring supervised topic modeling for depression-related language in Twitter. In Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality (pp. 99–107). (2015).
18. Cong, Q. et al. XA-BiLSTM: a deep learning approach for depression detection in imbalanced data. In 2018 IEEE international conference on bioinformatics and biomedicine (BIBM) (pp. 1624–1627). IEEE. (2018).
19. Priyamvada, B. et al. Stacked CNN-LSTM approach for prediction of suicidal ideation on social media. *Multimedia Tools Appl.* **82** (18), 27883–27904 (2023).
20. Jiang, Z. P., Levitan, S. I., Zomick, J. & Hirschberg, J. Detection of mental health from reddit via deep contextualized representations. In Proceedings of the 11th international workshop on health text mining and information analysis (pp. 147–156). (2020).
21. Nguyen, T., Yates, A., Zirikly, A., Desmet, B. & Cohan, A. Improving the Generalizability of Depression Detection by Leveraging Clinical Questionnaires. In 60th Annual Meeting of the Association for Computational Linguistic (pp. 8446–8459). ACL. (2022).
22. Lamichhane, B. Evaluation of Chatgpt for nlp-based mental health applications. *ArXiv Preprint arXiv* 230315727. (2023).
23. Ngo, B. H., Bui, D. C. & Choi, T. J. How to enrich cross-domain representations? Data augmentation, cycle-pseudo labeling, and category-aware graph learning. *Expert Syst. Appl.* **271**, 126597 (2025).
24. Garg, M., Liu, X., Sathvik, M. S. V. P. J., Raza, S. & Sohn, S. MultiWD: Multi-label wellness dimensions in social media posts. *J. Biomed. Inform.* **150**, 104586 (2024).
25. Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies*, volume 1 (long and short papers) (pp. 4171–4186). (2019).
26. Lan, Z. et al. Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942. (2019).
27. Sanh, V., Debut, L., Chaumond, J. & Wolf, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108. (2019).
28. Vajre, V., Naylor, M., Kamath, U. & Shehu, A. PsychBERT: a mental health language model for social media mental health behavioral analysis. In: *2021 IEEE International Conference on Bioinformatics and Biomedicine*(BIBM) (pp. 1077–1082). IEEE. (2021).
29. Ji, S. et al. Mentalbert: publicly available pretrained Language models for mental healthcare. *ArXiv Preprint arXiv* 211015621. (2021).
30. Huang, K., Altosaar, J. & Ranganath, R. Clinicalbert: Modeling clinical notes and predicting hospital readmission. arXiv preprint arXiv:1904.05342. (2019).

Author contributions

CY Hsieh supervised the research and reviewed the manuscript. QY Ye developed the method and implemented the code. FC Liu provided consultation support and reviewed the manuscript. X Wang engaged in reviewing the manuscript. CH Lee participated in the manuscript review and contributed relevant research concepts. CS Lin conducted the research and contributed to the writing of the manuscript.

Funding

This work is financially supported by Lumosa Therapeutics-Taipei Medical University A-114-078 and the National Science and Technology Council (NSTC) of Taiwan under Grant 114-2320-B-038-026 -.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to C.-S.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025