CURE: Context- and Uncertainty-Aware Mental Disorder Detection

Migyeong Kang[1] **Goun Choi**[1] **Hyolim Jeon**[1]
Jihyun An[2] **Daejin Choi**[3*] **Jinyoung Han**[1*]
[1]Department of Applied Artificial Intelligence, Sungkyunkwan University, South Korea
[2]Department of Psychiatry, Samsung Medical Center, South Korea
[3]Department of Computer Science & Engineering, Incheon National University, South Korea
jinyounghan@skku.edu, djchoi@inu.ac.kr
{gy77, gwcat0506, gyfla1512}@g.skku.edu, jh85.an@samsung.com

Abstract

As the explainability of mental disorder detection models has become important, symptom-based methods that predict disorders from identified symptoms have been widely utilized. However, since these approaches focused on the presence of symptoms, the context of symptoms can be often ignored, leading to missing important contextual information related to detecting mental disorders. Furthermore, the result of disorder detection can be vulnerable to errors that may occur in identifying symptoms. To address these issues, we propose a novel framework that detects mental disorders by leveraging symptoms and their context while mitigating potential errors in symptom identification. In this way, we propose to use large language models to effectively extract contextual information and introduce an uncertainty-aware decision fusion network that combines predictions of multiple models based on quantified uncertainty values. To evaluate the proposed method, we constructed a new Korean mental health dataset annotated by experts, named KoMOS. Experimental results demonstrate that the proposed model accurately detects mental disorders even in situations where symptom information is incomplete.

1 Introduction

Mental disorders have become an urgent global issue. The number of individuals suffering from mental disorders exceeds one billion, approximately 16 percent of the world's population (Rehm and Shield, 2019). Such severity has in turn leveraged the importance of detecting mental disorders, which determines whether an individual can be in a danger of mental illness or not. Social media has been considered as one of the key sources widely used in mental disorder detection research, as its anonymity property encourages individuals with mental illness to reveal their mental health status or self-disclosure rarely concerning about social stigma (De Choudhury and De, 2014; Tadesse et al., 2019; Kim et al., 2023). Connecting individuals to many peers who share similar experiences to mental health experts online without temporal and space restriction, social media has populated a deluge of data, which has been widely used for mental disorder detection (De Choudhury et al., 2013, 2016; Shen and Rudzicz, 2017).

The growing importance of mental disorder detection and the availability of abundant data have attracted research communities to develop diverse deep learning models for the mental disorder detection task (Yates et al., 2017; Chen et al., 2018; Dutta and De Choudhury, 2020; Lee et al., 2020). Unfortunately, despite their significant efforts in performance improvement, the black-box property of deep learning models can be a limitation in providing interpretation of the model results or evidences of a decision, which degrades the trustworthiness (Watson et al., 2019). Leveraging the importance of explainability of the model results in mental disorder detection, there have been a few attempts that find psychiatric symptoms in detect mental disorders (Zogan et al., 2022; Nguyen et al., 2022; Chen et al., 2023; Song et al., 2023). These approaches, denoted as *symptom-based* models, consist of a two-step pipeline in general: (i) a symptom identification from user-generated posts, which computes a vector representing the presence or likelihood of symptoms, and (ii) a disease detection to predict mental disorders using the calculated symptom vector. Note that only the symptom vector is used in the second step. Compared to post-based approaches, these approaches not only detect mental disorders with high performance, but also provide an interpretability of the model results as the symptom vector can be used in interpreting how the decision was made by the likelihood of the symptoms (Zhang et al., 2022; Chen et al., 2023; Song et al., 2023).

*Corresponding authors.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 17924–17940

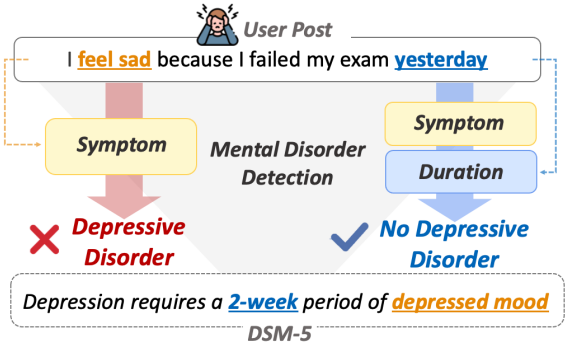

Figure 1: This example demonstrates the importance of contextual information in mental disorder detection. Unlike symptom-based approaches that rely solely on symptom information (left), accurate predictions are possible when contextual information (e.g., duration) is considered (right).

However, relying solely on a symptom vector in the *symptom-based* approaches can be limited in capturing context information around symptoms and diseases, resulting in inaccurate decision for mental disorder. According to the DSM-5 (Diagnostic and Statistical Manual of Mental Disorders), it is suggested to consider not only the indicated symptoms, but also contextual information of symptoms, including duration, frequency, and causes, for diagnosis. As Figure 1 illustrates, for example, a criterion for diagnosing major depressive disorder includes experiencing a depressed mood during the same *2-week period* (American Psychiatric Association et al., 2013). Thus, the post *"I feel sad because I failed my exam **yesterday**."* should be decided as non-depression while the post *"I have been feeling sad **for a month**."* indicates potential depression. Unfortunately, the symptom-based models are likely to determine both posts as depression disorder, since the symptom is explicitly indicated in the two posts. Furthermore, *uncertainty* of a symptom-based model in computing symptom vectors should be considered in detecting mental disorders. Since the symptom-based models depend on only symptom vectors calculated from user posts, a symptom vector with low confidence can result in inaccurate decision. Note that recent work has highlighted the risk of the current symptom identification models showing suboptimal accuracy due to difficulty in capturing complex symptom expression from user posts (Gupta et al., 2022; Zhang et al., 2022, 2023).

To address these limitations, we propose **CURE** (**C**ontext- and **U**ncertainty-aware Mental Diso**R**der D**E**tection), a novel approach for detecting mental disorders that (i) uses not only the presence or likelihood of symptoms, but also their contextual information and (ii) reduces the potential uncertainty originated from the symptom identification. To this end, the proposed model is designed to cooperates with a large language model (LLM) showing a capability in natural language understanding to effectively extract contextual factors pre-defined from the guidance of psychiatrists. Consisting of five different sub-models and the uncertainty-aware decision-fusion network with Spectral-normalized Neural Gaussian Process (SNGP) (Liu et al., 2023), the proposed model can detect mental disorders accurately even when symptoms are hardly captured from a given post. Our evaluation with the newly collected and annotated dataset named as KoMOS (Korean Mental Health Dataset with Mental Disorder and Symptoms labels), which contains 6,349 Q&A pairs created through the interaction between a user and an expert for mental health demonstrates that the proposed model outperforms the state-of-the-art models, showing robust performance even when symptom identification is incorrect. We summarize the contributions of this paper as follows.

- We propose **CURE**, a novel approach for detecting mental disorders, which can effectively capture both the symptom and their context information by cooperating with a large language model. It also employs an uncertainty-aware decision fusion network, which can enhance the model performance from potential problem of the incomplete or inaccurate symptom vectors.
- We build and publicly release KoMOS[1] (Korean Mental Disorder Dataset with Symptom), a question-answer dataset between a user and an expert in mental health. The dataset covers 4 mental disorders with 28 symptoms, which were labeled based on a guidance from professionals.

2 Related Work

2.1 Mental Health Detection for User Posts

A popular approach in detecting mental disorders from social media is to extract and use the features from a user's post to detect whether the writer is experiencing mental disorders or not (Gaur et al., 2019; Amini and Kosseim, 2020; Yoon et al., 2022). Although previous research has focused on single disorders (Yates et al., 2017; Chen et al.,

[1]https://github.com/gyeong707/EMNLP-2024-CURE

2018; Lee et al., 2022, 2024), recent studies have emerged in exploring the prediction of multiple disorders (Dinu and Moldovan, 2021; Kim et al., 2020; Chen et al., 2023). In these post-based approaches, pre-trained language models such as BERT (Devlin et al., 2018) have been widely used to capture linguistic features in social media posts (Murarka et al., 2020; Park et al., 2020; Dinu and Moldovan, 2021; Lee et al., 2023). Unfortunately, these approaches are less likely to provide interpretability of the model results to verify whether a model determines with suitable (medical) rationales, which is essential for mental disorder detection.

With the great success of large language models (LLMs) in various domains, there have been attempts to utilize LLMs for mental health detection (Yang et al., 2023, 2024; Xu et al., 2024). Focusing mainly on applicability of LLMs in mental health detection, these studies demonstrated that LLMs have the great capabilities in explaining the results of mental disorder detection, but their performance is lower than that by the supervised methods due to the lack the knowledge required to detect mental illnesses (Yang et al., 2024). Therefore, instead of using LLMs for direct prediction, we propose a way of interacting with an LLM to extract the context of symptoms from user-generated posts, by exploiting strong ability of LLMs in natural language understanding.

2.2 Symptom-based Methods for Mental Disorder Detection

To provide explainability, the recently proposed models have adopted diagnostic tools such as PHQ-9 (Kocalevent et al., 2013) and DSM-5 (American Psychiatric Association et al., 2013), which were designed with medical reasoning. For example, Nguyen et al. (2022) developed a model to find the presence of symptoms described in PHQ-9 for mental disorder detection. Similarly, Zhang et al. (2022) investigated DSM-5 to build a set of symptoms and developed a symptom-based detection model for seven mental disorders. To simplify the process of symptom-based approaches, Song et al. (2023) proposed an end-to-end model using a Siamese network (Koch et al., 2015), which can be flexibly adapted to detection tasks for different mental illnesses.

However, the models proposed in these studies rely only on the presence or probability of symptoms, which can lead to incorrect predictions of mental disorders. Chen et al. (2023) have attempted to address this issue by using the symptom vectors with a post embedding calculated by mental-bert (Ji et al., 2021), with an expectation that a post embedding may include contextual information of symptoms. Unfortunately, the contextual information in a post embedding is implicit, which are unlikely to provide explanations for what and how context contributes to mental disorder detection. Therefore, this study aims to explicitly integrate the context of symptoms into the model for detecting mental disorders, which not only improves the performance, but also allows not to lose explainability. Furthermore, we reduce the risk from the uncertainty of the model on symptom identification by adopting an uncertainty-aware fusion network.

3 Dataset Construction

In this section, we describe how to build **KoMOS**, a novel **Ko**rean **M**ental health dataset with mental dis**O**rder and **S**ymptoms labels.

3.1 Data Collection

We collected data from Naver Knowledge iN, a popular Q&A platform in Korea, where users can anonymously ask questions and receive answers on various topics. From the mental health category in Naver Knowledge iN, we first collected the posts that were answered by the certified psychiatrists to ensure the quality of our dataset, uploaded from October 2008 to September 2021. We then manually reviewed the collected Q&A pairs to filter out the pairs that neither the questions are not the request for a diagnosis nor the certified psychiatrists held off decisions of mental disorders due to the lack of information. Throughout the process, we finally obtain 8,000 Q&A pairs.

3.2 Data Annotation

For the collected Q&A pairs, we conduct the annotation process to label mental disorders and symptoms for each post under the supervision of psychiatrists.

Disorder Labeling. To assign a disorder label to a given post, we first extracted the disorder decision (from answers in Q&A pairs) written by the certified psychiatrists as the answer. We then selected the posts with four major disorders in the dataset: (i) Depressive Disorders, (ii) Anxiety Disorders, (iii) Sleep Disorders, and (iv) Eating Disorders. In addition, we considered the posts where an expert explicitly mentioned that the described status or physical reaction is either instantly caused

by an stress or not much far from an ordinary status, which are labelled as 'Non-Disease'. The total number of the selected posts are 6,349.

Symptom Labeling. Unlike disorders labeling, we design a new process to annotate symptoms as they are unlikely to be indicated explicitly in the answers. In particular, we carefully reviewed DSM-5 and extracted a list of symptoms for each disorder. The list is then refined by a psychiatrist, which resulted in a total of 28 symptom labels. Following the annotation criteria (established by psychiatrists) and considering the disorder label, we finally annotated a set of symptoms revealed in individual posts.

Psychiatrists Validation. We validated the annotated data with two additional psychiatrists. In particular, each psychiatrist independently labeled disorders and symptoms for 150 posts sampled from the annotated dataset, and then the agreements among the labels by two experts and ours were measured by Krippendorff's alpha, a statistical measure of inter-rater agreement. The agreement scores of disorder and symptom annotation are more than 90% and 82.9%, respectively, indicating that the annotation process works correctly. The detail of the validation results is presented in Appendix A.

3.3 Dataset Description

Our final dataset consists of 6,349 posts across the five mental disorder categories and their corresponding 28 symptoms. In detail, there are 1,408, 1,107, 1,783, 884, and 1,464 posts for depressive disorders, anxiety disorders, sleep disorders, eating disorders, and non-disease, respectively. Note that 286 posts show multiple disorders showing comorbidity, e.g., both depression and sleep disorder are observed in a post.

4 The Model

Defining the mental disorder detection task as a multi-label classification problem, the model takes a user's post as an input and predicts five labels (4 disease and non-disease) independently. In this section, we describe the proposed model in detail.

4.1 Overall Architecture

An overview of the proposed model is illustrated in Figure 2. Our model consists of the three main parts: **(i) Feature Extraction** that is responsible for extracting the necessary features from user posts, including symptom and contextual information, **(ii) Model Prediction** where sub-models trained on different combinations of features generate each prediction, and **(iii) Uncertainty-aware Decision Fusion** that calculates the uncertainty of the predictions based on SNGP (Liu et al., 2023) and generates a final decision by fusing each prediction with quantified uncertainties.

4.2 Feature Extraction

In this section, we extract the features needed for disorder detection from user posts.

4.2.1 Symptom Identification

We develop a symptom identification model to detect psychiatric symptoms in user posts. The model is a multi-label classifier that produces a likelihood for each of 28 symptoms. We employ BERT (Devlin et al., 2018), which is widely used for predicting symptoms due to its rich representations of text (Zhang et al., 2022; Nguyen et al., 2022). We fine-tune the model to identify symptoms and obtain the `[CLS]` token for each post, which is known as the aggregated representation of BERT. This representation is then passed through a fully connected layer to produce likelihood vectors. Given a user post p, the corresponding likelihood S is,

$$S = sigmoid(W \cdot \text{BERT}_{symp}(p) + b) \quad (1)$$

Subsequently, a predefined threshold is applied to transform these likelihoods into the final predictions.

4.2.2 Context Extraction

To improve performance of disorder detection, we additionally utilize contextual information of symptoms revealed in posts. Since each post inherently includes relevant contextual information, such as duration or frequency of symptom, embeddings of the encoded post can be utilized as contextual information. However, we found that the post-based approach leveraging encoded posts as input tends to overly rely on representative keywords that prominently appear in data related to disease, leading to insufficient consideration of the symptom context. Motivated by this, our objective is to extract contextual information of each symptom from posts and explicitly integrate it into the model.

To this end, we first define eight crucial context factors under a psychiatrist's guidance: *Cause*, *Frequency*, *Duration*, *Age*, and four types of Affects - *Social*, *Academic*, *Occupational*, and *Life-threatening*. We then create instructions with descriptions for each factor, and request a large language model (LLM) to identify these factors from

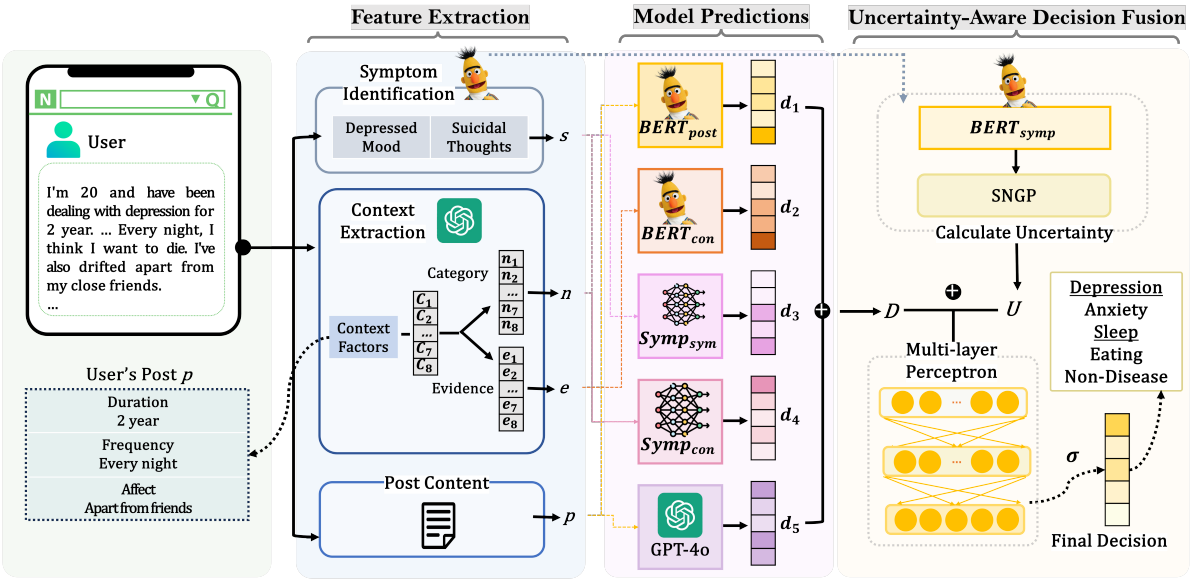

Figure 2: **CURE** consists of three procedures: *(1) Feature Extraction*, which involves extracting symptoms and contextual information from user posts; *(2) Model Predictions*, where sub-models trained by different combinations of features generate individual predictions; and *(3) Uncertainty-Aware Decision Fusion*, which generates the final decision based on a quantified uncertainty.

user posts. Contextual information is extracted in two ways: (i) `category` as a numerical value classified by defined criteria *(e.g., 0 for symptoms less than a month, 1 for more than a month)* and (ii) `evidence` as a direct expression for the context in text *(e.g., for two years)*. These two types of context are subsequently utilized in different ways in detecting mental disorders. Additionally, we obtain symptoms revealed in user posts from a large language model, which are not used as contextual factors. The details including definitions of factors and prompt designs can be found in Appendix B. The context factors, $C_i = \{c_1, \ldots, c_8\}$, are extracted as follows:

$$n_i = \text{GPT-4o}_{category}(p, c_i) \quad (2)$$

$$e_i = \text{GPT-4o}_{evidence}(p, c_i) \quad (3)$$

where n_i represents the numerical value for a context `category`, and $N = \{n_1; \ldots; n_8\}$ represents their concatenation. Similarly, e_i stands for context `evidence`, and $E = \{e_1; \ldots; e_8\}$ denotes the concatenated vector of them.

4.3 Model Predictions

The straightforward approach to consider both symptoms and contextual information is training the model using them as input feature. However, combining features of inherently different natures can result in sub-optimal performance, e.g., by interfering with each other. For example, a model training with different modalities under identical settings, such as the same learning rate and the number of iterations, can suffer from gradient vanishing in specific modalities (Wang et al., 2020; Yao and Mihalcea, 2022). To address this issue, we build multiple sub-models, where different modalities are used as input, and then fuse their predictions, which allows to leverage the strengths of multiple feature with different natures.

Consequently, we develop five *sub-models* that utilize different combinations of three features: *symptoms*, *context*, and *posts*. Here, we employ BERT (Devlin et al., 2018) and Symp (Zhang et al., 2022) as backbone model. The considered sub-models are summarized as follows: (i) BERT_{post} that utilizes only the post content p, (ii) $\text{BERT}_{context}$ that uses only context `evidence` E, along with symptoms extracted from the LLM (iii) $\text{Symp}_{symptom}$ that leverages only symptom vector S, and (iv) $\text{Symp}_{context}$ considering both symptom and context `category` as input by concatenating two vectors $\{S; N\}$. Additionally, we consider a fifth sub-model, a GPT-4o, which utilizes the user's post p as the input. We train each model based on the defined features, and then obtain prediction results $D = \{d_1, \ldots, d_5\}$ from these models.

4.4 Uncertainty-Aware Decision Fusion

To fuse the predictions of the sub-models for final decision-making, we introduce the uncertainty-aware decision fusion network. To this end, we first

quantify the uncertainty of each prediction. The uncertainty reflects the confidence of the model prediction. Therefore, the influence of models with high uncertainty should be reduced during the final decision-making process. This approach helps us make robust decisions, preventing errors in symptom identification.

To calculate the uncertainty of prediction, we utilize a Spectral-normalized Neural Gaussian Process (SNGP) (Liu et al., 2023), which estimates uncertainty for each prediction in terms of out-of-distribution probability. This is based on the fact that a trained model might produce incorrect decisions when the input data significantly differs from the distribution of training data. In our study, we estimate the uncertainty of BERT$_{symp}$, which is the model for symptom identification, and its uncertainty reflects the confidence in the predicted symptoms. As a result, the uncertainty obtained from the models is denoted as U.

Finally, we fuse the estimated uncertainties U and all predictions D of the sub-models to make a final decision. After that, the data is passed through an MLP layer, fusing the predictions of all sub-models with their uncertainty values. This can be expressed as follows:

$$H = \sigma(W_1 \cdot CONCAT(U, D) + b_1) \quad (4)$$

$$\hat{y} = sigmoid(W_3 \cdot \sigma(W_2 \cdot H + b_2) + b_3) \quad (5)$$

The objective of the training process is to minimize the binary cross-entropy loss between the predicted label $\bar{y}$ and the target y.

5 Experimental Results

5.1 Baselines

We consider two types of approaches as baselines for extensive performance comparison.

Post-based Approach. SVM+TF-IDF (Abd Rahman et al., 2020) is a traditional machine learning method using linguistic features. HAN (Sekulić and Strube, 2020) hierarchically encodes text with word and sentence-level attention. BERT (Devlin et al., 2018) is pre-trained model that is widely used for disorder detection (Murarka et al., 2020).

Symptom-based Approach. PHQ-9 (Nguyen et al., 2022) utilizes a hidden state obtained from symptom identification model to preserve semantic information. Symp (Zhang et al., 2022) leverages the likelihood vector of symptoms. PsyEx (Chen et al., 2023) has a two-stream architecture considering the likelihood of symptoms and post-embedding.

5.2 Evaluation Metrics

To assess the performance of the proposed method, we employ two metrics: Recall and F1 Score. In the context of mental disorder detection, minimizing False Negatives is crucial to reduce the risk of misdiagnosis. In this context, we select Recall as an important indicator of model prediction performance. However, models with high recall may also have high False Positive rates, risking misclassification of healthy individuals as having an illness, leading to potential overdiagnosis. Since our research goal is to develop a model that minimizes overdiagnosis while maintaining high Recall, we introduce the F1 Score as an additional evaluation metric.

5.3 Experimental Setting

We conduct experiments on detecting disorders in multi-label scenarios. The entire dataset is divided into five folds with an 8:2 ratio, then the average performance of all folds is reported. For training sub-models, we utilize the BERT model pre-trained on the Korean corpus, `klue/bert-base` and the Symp (Zhang et al., 2022). Note that the validation set remains consistent throughout our entire process. In the final decision making, we use $lr = 0.003$, batch size $= 64$, and optimizer $=AdamW$. For additional settings of experiments, please refer to Appendix C.

5.4 Evaluation on Feature Extraction

We first evaluate how the features are accurately extracted.

Symptom Identification. We evaluate the results of symptom identification using BERT-based models. The detailed results are presented in Appendix D. The performance of both BERT-based models was comparable, with an average F1 score of 0.80. Nevertheless, the class imbalance led to a significant drop in performance for the underrepresented classes, emphasizing the need for an approach that can mitigate errors in symptom identification.

Context Extraction. To evaluate the context factors, 300 samples were randomly selected and manually evaluated by two annotators. Disagreements were resolved through discussion. Most factors, except for *Cause*, achieved over 93% accuracy, indicating effective contextual extraction by LLMs. Extracting the cause of psychiatric symptoms, often requires expertise, resulting in 84% accuracy. A detailed result is shown in Appendix B.

Feature	Model	Depressive Disorders		Anxiety Disorders		Sleep Disorders		Eating Disorders		Non-Disease	
		Rec.	F1.	Rec.	F1.	Rec.	F1.	Rec.	F1.	Rec.	F1.
Post	SVM+TF-IDF	0.637	0.731	0.706	0.776	0.696	0.771	0.850	0.864	0.175	0.291
	HAN	0.731	0.767	0.755	0.792	0.816	0.805	0.899	0.890	0.409	0.509
	BERT	0.820	0.813	0.851	0.840	**0.895**	0.840	0.947	0.911	0.482	0.601
Symptom	PHQ-9	0.810	0.801	0.838	0.831	0.836	0.817	0.889	0.892	0.415	0.542
	Symp	**0.827**	0.805	0.877	0.845	0.849	0.822	0.935	0.905	0.463	0.583
Symptom, Post	PsyEx	0.774	0.803	0.866	0.841	0.755	0.793	0.923	0.910	**0.606**	0.626
Symptom, Context Factor	$\text{BERT}_{context}$	0.802	0.803	0.831	0.828	0.867	0.829	0.940	0.913	0.488	0.582
	$\text{Symp}_{context}$	0.812	0.807	**0.891**	0.845	0.863	0.829	0.933	0.905	0.484	0.595
Fusion	**CURE**	0.825	**0.824**	0.875	**0.855**	0.874	**0.852**	**0.950**	**0.919**	0.576	**0.649**

Table 1: Results of mental disorder detection on the KoMOS dataset. Here, the first column represents the features utilized in the model. The average Recall and F1-Score from 5-fold cross-validation are reported. The highest results for each label are shown in bold, and the second-highest results are underlined.

Model	Disease (Avg)			Non-disease		
	Pre.	Rec.	F1.	Pre.	Rec.	F1.
GPT-3.5	0.673	0.916	0.771	0.717	0.235	0.353
GPT-4o	0.631	**0.977**	0.764	**0.840**	0.310	0.453
MentalLLaMa	0.268	0.748	0.369	0.290	0.453	0.353
CURE	**0.845**	0.881	**0.862**	0.745	**0.576**	**0.649**

Table 2: Performance of Large Language Models. We reported the average scores for the four diseases.

Model	Total (Avg)		Disease (Avg)		Non-disease	
	Rec.	F1.	Rec.	F1.	Rec.	F1.
CURE	**0.820**	**0.820**	**0.881**	**0.862**	**0.576**	**0.649**
w/o context	0.813	0.818	0.873	0.862	0.572	0.644
w/o uncertainty	0.813	0.819	0.876	0.862	0.561	0.647
w/o uncertainty+context	0.815	0.817	0.880	0.860	0.554	0.642

Table 3: Ablation study to examine the effectiveness of context information and uncertainty.

5.5 Evaluation on Mental Disorder Detection

We conduct a comprehensive evaluation to assess the effectiveness of the proposed framework for mental disorder detection.

5.5.1 Overall Performance

Table 1 presents the average recall and F1 scores for each of the four diseases and the non-disease category. In our dataset, it is crucial to accurately differentiate between diseases and non-disease cases while effectively capturing the risk of diseases. Our proposed method outperformed the baselines by achieving the highest F1 score and maintaining high recalls across all categories. Specifically, models using a single feature showed high sensitivity in predicting diseases, but they struggled to distinguish between disease and non-disease cases, resulting in lower performance in the non-disease category. On the other hand, PsyEx, which utilizes both symptom and post feature, showed strong detection capabilities for non-disease by leveraging their rich information. However, it showed a limitation in detecting diseases especially for depressive disorder and sleep disorder. We also reported results for two sub-models that utilized contextual features. While these models classified posts into disease categories from different perspectives, they did not improve overall performance. In contrast, our approach achieved consistent performance improvements by leveraging the strengths of sub-models trained on diverse features.

5.5.2 Performance on LLMs

We additionally examined the performance of large language models (LLMs) for detecting mental disorders. Table 2 shows the results of GPT-3.5 (gpt-3.5-turbo), GPT-4o (gpt-4o-2024-05-13), and MentalLLaMA (MentalLLaMA-chat-7B) (Yang et al., 2024) that is a fine-tuned version of LLaMa2 (Touvron et al., 2023) for various mental health detection tasks. The result demonstrates LLMs tend to over-diagnose in mental disorder detection, showing high performance in disease categories but poor performance in non-disease cases. This suggests that although LLMs can identify information related to mental disorders, they lack the expertise required to make accurate diagnostic decisions.

Notably, MentalLLaMA, which was optimized for mental health datasets, showed significantly lower performance compared to other GPT-based models. There could be a few reasons for this performance degradation. First, we used the smallest version of MentalLLaMA due to computational resource limitations. As this is much smaller than other GPT models, the lower natural language capabilities may affect performance. Second, MentalLLaMA only works for English-written text while other LLMs support multiple languages. To run MentalLLaMA, we had to translate our data into English using DeepL API[2], one of the translation tools, and then feed them into MentalLLaMA. Loss or distortion of context information during transla-

[2]https://www.deepl.com/

tion may also have contributed to the performance degradation.

In contrast, our approach demonstrated the ability to effectively distinguish between non-disease and disease categories, achieving the highest F1 scores. Detailed performance results for all categories are provided in Appendix E.

5.5.3 Ablation Study

To evaluate the contribution of each part in our proposed model, We performed an ablation study.

Analysis on Model Components. We performed an analysis to evaluate the performance of the main components of the model. First, we evaluated it after removing two sub-models that use contextual information. Then, we eliminated the uncertainty, in which case the model makes the final decision without the uncertainty of symptom identification. Finally, we assessed the performance after removing both components. As shown in Table 3, the performance declined when each component was removed, particularly affecting the recall across all categories. This suggests that contextual information plays a vital role in accurately identifying both disease and non-disease cases, and that uncertainty can be effectively leveraged in the final decision-making process, leading to improved overall performance.

Analysis on Sub-Models. To evaluate the impact of the sub-models utilized in our approach, we conducted an analysis on sub-models. To this end, we removed each of the five sub-models employed in our model and assessed the performance. The results in Table 4 demonstrate that all five proposed sub-models contribute to enhancing the final model performance. In particular, the BERT$_{post}$ and BERT$_{context}$ models were crucial for detecting non-disease cases. This indicates that the contextual information in post content is important to distinguish between disease and non-disease cases. Furthermore, removing the Symp$_{symptom}$ model showed the slightly improved performance for non-disease cases while the performance for disease categories is degraded. This result indicates that the model solely relying on symptom information plays more roles in predicting individual diseases and less contributes to identifying non-disease cases.

5.5.4 Analysis on Context Factors

We performed an additional analysis to assess how effective each context factor (*Age*, *Duration*, *Frequency*, and *Affect*) is in our model. To this end, we removed each context factor from the Symp$_{context}$ model, which uses both symptom and context information. In Table 5, we show that there was a decrease in performance across all cases, indicating the significant role of context factors. Detailed performance about the impact of context factors on each disease category is provided in Appendix E.

Model	Total (Avg)		Disease (Avg)		Non-disease	
	Rec.	F1.	Rec.	F1.	Rec.	F1.
CURE	**0.820**	**0.820**	**0.881**	**0.862**	0.576	0.649
w/o BERT$_{post}$	0.798	0.798	0.869	0.855	0.516	0.623
w/o BERT$_{context}$	0.808	0.816	0.872	0.859	0.553	0.643
w/o Symp$_{symptom}$	0.815	0.820	0.874	0.861	**0.579**	**0.655**
w/o Symp$_{context}$	0.810	0.819	0.869	0.862	0.573	0.647
w/o GPT-4o	0.810	0.815	0.872	0.858	0.562	0.643

Table 4: Ablation study to evaluate the impact of the five sub-models introduced in §4 Model Predictions.

Model	Total (Avg)		Disease (Avg)		Non-disease	
	Rec.	F1.	Rec.	F1.	Rec.	F1.
Symp$_{context}$	**0.797**	**0.796**	**0.875**	**0.846**	**0.484**	**0.595**
w/o duration	0.790	0.793	0.867	0.843	0.479	0.593
w/o age	0.784	0.792	0.861	0.844	0.479	0.588
w/o cause	0.784	0.792	0.864	0.844	0.465	0.583
w/o frequency	0.786	0.792	0.867	0.844	0.463	0.582
w/o affect	0.783	0.792	0.863	0.845	0.465	0.582

Table 5: Evaluation to investigate the impact of individual contextual factors on mental disorder detection.

5.5.5 Case Study

We conducted a case study to demonstrate how the proposed method can improve the detection of mental disorders. The prediction results for two representative cases are shown in Figure 3. In the first case, the writer expressed curiosity about having anorexia. The expert diagnosis indicated the writer did not have the disease, as no related symptoms were exhibited. Although the writer mentioned skipping meals due to lack of company, the symptom identification model inferred *"Difficulty in Eating"*, leading to an incorrect prediction. In contrast, the context-based model accurately predicted the writer did not have anorexia, demonstrating the effectiveness of our proposed method.

In the second case, the writer reported persistent chest discomfort since last year. The symptom identification model correctly inferred *"Chest Discomfort"*. However, it did not detect other symptoms typically associated with anxiety disorders. As a result, the symptom-based model did not classify this as a disease. In contrast, the context-based model recognized the duration of the symptom and accurately diagnosed it as an anxiety disorder, helping for final model to make an accurate prediction.

These two cases illustrate potential errors that can arise when relying solely on symptom information for disease detection. In both instances, our model can generate accurate predictions by uti-

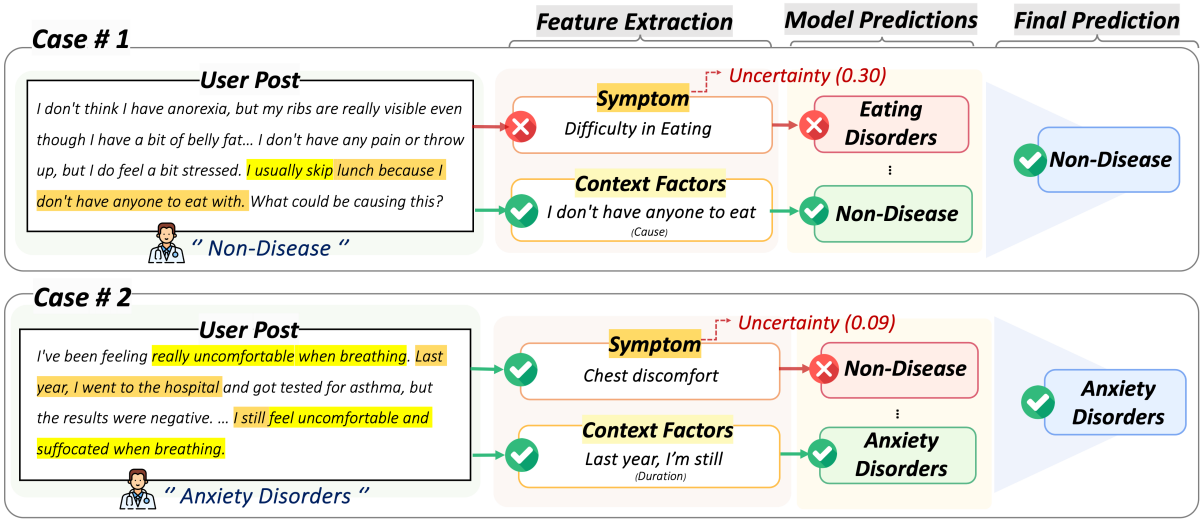

Figure 3: Prediction results for two cases in a case study. The results showcase three different approaches: $\text{Symp}_{symptom}$ focuses solely on symptoms, while $\text{BERT}_{context}$ incorporates contextual information. **CURE** is an ensemble model with the uncertainty-aware decision-fusion network.

lizing contextual information and an uncertainty-based decision fusion framework.

6 Conclusion

In this paper, we proposed a novel approach in mental disorders detection. Our model uses contextual information for more accurate detection and employs an uncertainty-aware decision fusion network to reduce errors. Additionally, we created KoMOS, the first mental health dataset in Korea, which will be publicly available to authorized researchers. Experiments show our approach outperforms existing methods. We believe the proposed model can provide early intervention and support by quickly detecting mental disorders on social media.

Limitations

This research has several limitations that can be addressed in future studies. Firstly, our study only addresses four specific mental disorders. Future work should aim to expand the scope to include a wider variety of mental health conditions to enhance the applicability and robustness of the proposed model. Secondly, we were unable to conduct experiments to test our model to different datasets. This limitation stemmed from the differences in data characteristics between our dataset and most publicly available ones. Our data was gathered from users who reported their conditions for diagnostic purposes, whereas previous research datasets comprised social media posts with various purposes such as emotion or opinion sharing (Yates et al., 2017; Cohan et al., 2018; Cai et al., 2023; Chen et al., 2023; Zhang et al., 2022). This discrepancy made it difficult to extract contextual factors we defined like duration, cause, and affects from these datasets. As a result, our model is well suited to self-reported data for diagnostic purposes, and it remains a goal of future research to define contextual factors that can be effectively leveraged from general social media datasets and how to exploit them.

Ethics Statement

We carefully considered and addressed any potential ethical issues that could arise during our research process. Specifically, we made sure not to collect any meta information, such as user IDs or nicknames, during the data construction phase. Additionally, during the annotation process, we removed any information that could identify the user, such as email addresses or affiliations. Therefore, we confirm that there are no ethical issues associated with the public use of the data used in this study. This study has been approved by the Institutional Review Board (SKKU 2024-10-001).

Acknowledgements

This research was supported by the MSIT (Ministry of Science, ICT), Korea, under the Global Research Support Program in the Digital Field program)(RS-2024-00425354) and Global Scholars Invitation Program (RS-2024-00459638) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

References

Rohizah Abd Rahman, Khairuddin Omar, Shahrul Azman Mohd Noah, Mohd Shahrul Nizam Mohd Danuri, and Mohammed Ali Al-Garadi. 2020. Application of machine learning methods in mental health detection: a systematic review. *Ieee Access*, 8:183952–183964.

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

Hessam Amini and Leila Kosseim. 2020. Towards explainability in using deep learning for the detection of anorexia in social media. In *Natural Language Processing and Information Systems: 25th International Conference on Applications of Natural Language to Information Systems, NLDB 2020, Saarbrücken, Germany, June 24–26, 2020, Proceedings 25*, pages 225–235. Springer.

Yicheng Cai, Haizhou Wang, Huali Ye, Yanwen Jin, and Wei Gao. 2023. Depression detection on online social network with multivariate time series feature of user depressive symptoms. *Expert Systems with Applications*, 217:119538.

Siyuan Chen, Zhiling Zhang, Mengyue Wu, and Kenny Zhu. 2023. Detection of multiple mental disorders from social media with two-stream psychiatric experts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9071–9084.

Xuetong Chen, Martin D Sykora, Thomas W Jackson, and Suzanne Elayan. 2018. What about mood swings: Identifying depression on twitter with temporal measures of emotions. In *Companion Proceedings of the The Web Conference 2018*, pages 1653–1660.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. 2015. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018. Smhd: a large-scale resource for exploring online language usage for multiple mental health conditions. *arXiv preprint arXiv:1806.05258*.

Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 71–80.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the international AAAI conference on web and social media*, volume 7, pages 128–137.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2098–2110.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Anca Dinu and Andreea-Codrina Moldovan. 2021. Automatic detection and classification of mental illnesses from general social media texts. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 358–366.

Sarmistha Dutta and Munmun De Choudhury. 2020. Characterizing anxiety disorders with online social and interactional networks. In *International Conference on Human-Computer Interaction*, pages 249–264. Springer.

Manas Gaur, Amanuel Alambo, Joy Prakash Sain, Ugur Kursuncu, Krishnaprasad Thirunarayan, Ramakanth Kavuluru, Amit Sheth, Randy Welton, and Jyotishman Pathak. 2019. Knowledge-aware assessment of severity of suicide risk for early intervention. In *The world wide web conference*, pages 514–525.

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. *arXiv preprint arXiv:2205.13884*.

Shaoxiong Ji, Tianlin Zhang, Luna Ansari, Jie Fu, Prayag Tiwari, and Erik Cambria. 2021. Mentalbert: Publicly available pretrained language models for mental healthcare. *arXiv preprint arXiv:2110.15621*.

Jina Kim, Jieon Lee, Eunil Park, and Jinyoung Han. 2020. A deep learning model for detecting mental illness from user content on social media. *Scientific reports*, 10(1):1–6.

Meeyun Kim, Koustuv Saha, Munmun De Choudhury, and Daejin Choi. 2023. Supporters first: Understanding online social support on mental health from a supporter perspective. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW1).

Rüya-Daniela Kocalevent, Andreas Hinz, and Elmar Brähler. 2013. Standardization of the depression screener patient health questionnaire (phq-9) in the general population. *General hospital psychiatry*, 35(5):551–555.

Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille.

Daeun Lee, Hyolim Jeon, Sejung Son, Chaewon Park, Ji hyun An, Seungbae Kim, and Jinyoung Han. 2024. Detecting bipolar disorder from misdiagnosed major depressive disorder with mood-aware multi-task learning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4954–4970.

Daeun Lee, Migyeong Kang, Minji Kim, and Jinyoung Han. 2022. Detecting suicidality with a contextual graph neural network. In *Proceedings of the eighth workshop on computational linguistics and clinical psychology*, pages 116–125.

Daeun Lee, Soyoung Park, Jiwon Kang, Daejin Choi, and Jinyoung Han. 2020. Cross-lingual suicidal-oriented word embedding toward suicide prevention. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2208–2217.

Daeun Lee, Sejung Son, Hyolim Jeon, Seungbae Kim, and Jinyoung Han. 2023. Towards suicide prevention from bipolar disorder with temporal symptom-aware multitask learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4357–4369.

Jeremiah Zhe Liu, Shreyas Padhy, Jie Ren, Zi Lin, Yeming Wen, Ghassen Jerfel, Zachary Nado, Jasper Snoek, Dustin Tran, and Balaji Lakshminarayanan. 2023. A simple approach to improve single-model deep uncertainty via distance-awareness. *Journal of Machine Learning Research*, 24(42):1–63.

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

Thong Nguyen, Andrew Yates, Ayah Zirikly, Bart Desmet, and Arman Cohan. 2022. Improving the generalizability of depression detection by leveraging clinical questionnaires. *arXiv preprint arXiv:2204.10432*.

Sungjoon Park, Kiwoong Park, Jaimeen Ahn, and Alice Oh. 2020. Suicidal risk detection for military personnel. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2523–2531.

Jürgen Rehm and Kevin D Shield. 2019. Global burden of disease and the impact of mental and addictive disorders. *Current psychiatry reports*, 21(2):1–7.

Ivan Sekulić and Michael Strube. 2020. Adapting deep learning methods for mental health prediction on social media. *arXiv preprint arXiv:2003.07634*.

Judy Hanwen Shen and Frank Rudzicz. 2017. Detecting anxiety through reddit. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality*, pages 58–65.

Hoyun Song, Jisu Shin, Huije Lee, and Jong C. Park. 2023. A simple and flexible modeling for mental disorder detection by learning from clinical questionnaires. In *Annual Meeting of the Association for Computational Linguistics*.

Michael M Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2019. Detection of depression-related posts in reddit social media forum. *IEEE Access*, 7:44883–44893.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Weiyao Wang, Du Tran, and Matt Feiszli. 2020. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12695–12705.

David S Watson, Jenny Krutzinna, Ian N Bruce, Christopher EM Griffiths, Iain B McInnes, Michael R Barnes, and Luciano Floridi. 2019. Clinical applications of machine learning algorithms: beyond the black box. *Bmj*, 364.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. 2024. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–32.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Mentallama: Interpretable mental health analysis on social media with large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 4489–4500.

Yiqun Yao and Rada Mihalcea. 2022. Modality-specific learning rates for effective multimodal additive late-fusion. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1824–1834.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. *arXiv preprint arXiv:1709.01848*.

Jeewoo Yoon, Chaewon Kang, Seungbae Kim, and Jinyoung Han. 2022. D-vlog: Multimodal vlog dataset for depression detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12226–12234.

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.

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.

Hamad Zogan, Imran Razzak, Xianzhi Wang, Shoaib Jameel, and Guandong Xu. 2022. Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. *World Wide Web*, 25(1):281–304.

A Details of KoMOS

This section introduces the details of the KoMOS dataset, including data example, representative symptoms for each disorder, and expert validation.

A.1 Example of KoMOS

KoMOS consists of user-generated questions that provide detailed descriptions of their mental state and certified psychiatrists' diagnostic responses. An example from KoMOS dataset can be found in Table 8.

A.2 List of Representative Symptoms

We reviewed the DSM-5 to extract representative symptoms for each disorder. We found that anxiety disorders had the most representative symptoms, and the symptom 'loss of appetite' was representative of both depression and eating disorders. The detailed information is presented in Table 9.

A.3 Results of Expert Validation

In our dataset, diseases were labeled by collecting expert answers to user questions from the Naver Knowledge iN website, while symptoms were labeled based on annotation criteria established in collaboration with psychiatrists. To evaluate the quality of the labeled data, we randomly selected 150 samples and measured the agreement scores between our labels and those of two psychiatrists for both disorders and symptoms. The results showed an average agreement score of 92% for disorders and 83% for symptoms. Detailed results for disease and symptom annotations are presented in Tables 6 and 7, respectively.

Agreement Score for Disorders			
	Psychiatrist 1	Psychiatrist 2	Ours
Psychiatrist 1	1.0	-	-
Psychiatrist 2	0.985	1.0	-
Ours	**0.917**	**0.932**	**1.0**

Table 6: Psychiatrists validation on disorder labels.

Agreement Score for Symptoms			
	Psychiatrist 1	Psychiatrist 2	Ours
Psychiatrist 1	1.0	-	-
Psychiatrist 2	0.952	1.0	-
Ours	**0.855**	**0.804**	**1.0**

Table 7: Psychiatrists validation on symptom labels.

B Details of Context Extraction

In this section, we introduces the details of context extraction, including the meanings of each context factors, the procedure for extracting context factors from the large language model, and the prompt template.

B.1 Meaning of Context Factors

We extracted the eight contextual factors from user-generated post by using GPT-4o which is one of the large language models introduced by OpenAI. Each factor represents the following:

- *Cause* indicates whether there is an event or trigger that led the user to experience psychiatric symptoms. It is labeled as 0 if not mentioned, and 1 if mentioned.
- *Duration* indicates the period during which the user experienced psychiatric symptoms. It is labeled as 0 if not mentioned, 1 if less than one month, and 2 if more than one month.
- *Age* indicates the user's age. It is labeled as 0 if not mentioned, 1 if the user is a minor, and 2 if the user is an adult.
- *Frequency* indicates how often the user experiences psychiatric symptoms. It is labeled as 0 if not mentioned, 1 if less than three times a week, and 2 if more than three times a week.
- *Affect* is composed of four categories, indicating the impact of user's psychiatric symptoms affect *social*, *educational*, *occupational*, or *life-threatening* functioning.

B.2 Procedure of Extracting Context Factors

We use LLM to extract the eight context factors defined under the guidance of experts. The detailed procedure for context extraction is as follows:

	Data Example
User Question	Since last week, whenever I go out, I feel tightness in my chest and have difficulty breathing. When I'm in crowded places like the subway station, I also feel a bit dizzy. I thought it might just be my imagination, so I felt hesitant about going to the hospital for this. But today, while working part-time, it seems to have become more difficult to bear. I've received treatment for depression in the past, so I'm wondering if it could be panic disorder, but apart from feeling suffocated, there aren't any clear symptoms, so I'm a bit uncertain. I'm questioning whether these symptoms warrant a visit to the hospital.
Expert Answer	Hello. I'm Dr. Kim, a psychiatrist with the Korean Medical Association. The symptoms you mentioned can be mood and physical symptoms that can accompany severe anxiety. You mentioned that you had been treated for depression before; if your depression symptoms are not well-controlled recently, anxiety can also accompany them. In the case of panic disorder, symptoms such as palpitations, difficulty breathing, fear of dying, dizziness, etc., appear very intensely over 30 minutes to an hour and are accompanied by anticipatory anxiety about the possibility of another attack. While it is possible in your case, it might be a mild manifestation of agoraphobia among anxiety disorders rather than panic disorder. If the symptoms gradually worsen or affect your daily life, it is recommended to visit a hospital.

Table 8: An example of a KoMOS dataset. In our dataset, each data instance consists of a pair: a question from a user discussing their mental health problem, and an answer from a psychiatrist providing a diagnosis.

	Representative Symptoms
Depressive Disorders	Depressed mood, Irritability, Suicidal ideation, Self-harm, Loss of energy, Loss of appetite
Sleep Disorders	Hypersomnia, Insomnia, Sensory disturbances during sleep, Paresthesia, Sleep paralysis, Poor quality of sleep, Parasomnias, Excessive sleepiness
Eating Disorders	Dietary restrictions, Difficulty in eating, Self-induced vomiting, Fear of weight gain, Binge eating episodes, Loss of appetite
Anxiety Disorders	Palpitations, Chest discomfort, Tremor, Sweating, Abdominal discomfort, Feeling anxious, Other symptoms due to anxiety, Social anxiety, Phobias

Table 9: A list of representative symptoms for each mental disease. This list was selected with psychiatrist guidance based on the DSM-5.

1. **Selection of context factors**: The selection of context factors was primarily based on the DSM-5. For example, as the diagnostic criteria for major depressive disorder in DSM-5 include "experiencing a depressed mood during the same 2-week period," we can select duration as one of the context factors. After the selection process, candidates of context factors are then reviewed by experts to finalize the selection.

2. **Prompt initialization**: Next, we constructed prompts to extract each factor from posts using the LLM. The prompts include a brief description of each factor and 2-3 examples to guide the extraction. The initial prompt was designed to extract all eight factors from a single prompt, but this often failed due to the high complexity of the task and the inability to provide sufficient examples covering all factors.

3. **Prompt improvement**: Since constructing individual prompts for each factor requires a high cost for API, we needed to create prompts that could extract high-quality factors with minimal cost. To achieve this, we conducted an iterative improvement process by applying various prompting strategies (e.g., Chain-of-Thought), to a few samples from the training dataset. As a result, we finalized four prompts that were most effective for our purposes.

4. **Quality Evaluation**:To evaluate the efficacy of the final prompts in extracting contextual factors, we selected 300 sample data instances from the training set. Two annotators performed qualitative evaluations for each factor, and as a result, most factors showed 93% accuracy, except for Cause. The Cause factor, which indicates the cause for the symptoms, often requires more complex reasoning for extraction. For example, for a post, *"I go to bed at 2 AM and wake up at 6 AM, and I'm too sleepy during the day,"* the LLM has to infer that *"lack of sleep"* is the cause.

Factor	**Acc.**	**Factor(Affect)**	**Acc.**
Cause	0.84	Social	0.96
Frequency	0.96	Occupational	0.96
Duration	0.93	Educational	0.95
Age	0.95	Life-threatening	0.93

Table 10: Quality evaluation for each context factor extracted from large language model.

B.3 Prompt Templates of Context Extraction

Designed using the Chain-of-thought methodology (Wei et al., 2022), the prompts enable the LLM to extract factors through intermediate reasoning steps. Each of prompts aims to extract the Cause (Figure 4), Frequency (Figure 5), Age and Duration (Figure 6), and four Affects (Figure 7), respectively.

B.4 Quality Validation for Context Factors

To evaluate the LLM's contextual reasoning ability, we performed a quality assessment of the eight

Cause Extraction Prompt

I will show you a text in which an anonymous user describes their mental status. Based on this, add labels according to the following criteria:

Cause: Has the user experienced an event that could cause stress?
1: Yes. The user has experienced an event that could cause stress.
0: No. There is no indication of a stress-related event experienced by the user.

Let's think step by step.
1. Identify the main symptoms the user is experiencing.
2. Identify any stressful events that might be causing those symptoms.

If you output 1, you must provide clear evidence from the text. Output 0 if there are no mentions, and "No mention" for the reason.

EXAMPLE
Post: "...."
Answer: "...."

Figure 4: Prompt template of Cause

Frequency Extraction Prompt

I will show you a text in which an anonymous user describes their mental status. Based on this, add labels according to the following criteria:

Frequency: How often do users experience psychiatric symptoms?
2: Symptoms are expressed frequently (3 or more times per week) or every day.
1: Symptoms are expressed Intermittently (less than 3 times a week) or only in specific situations.
0: The frequency is not mentioned in the given text.

Let's think step by step.
1. Identify the main symptoms the user is experiencing.
2. Determine how often the user experiences these symptoms.

If you output 1 or 2, you must provide clear evidence. Output 0 if there are no mentions, and "No mention" for the reason.

EXAMPLE
Post: "...."
Answer: "...."

Figure 5: Prompt template of Frequency

extracted factors. We randomly sampled 300 data points, and two annotators evaluated the accuracy of the extracted factors. The results showed that most factors, except for cause, achieved an accuracy of over 90%, indicating that LLM's contextual inference capabilities can effectively extract important context from mental health posts. The results are shown in Table 10.

C Experimental Settings

C.1 Data Split Settings

We used 80% of the total 6,349 data, i.e., 5,076 cases, as training data for symptom identification and disorder detection tasks. The remaining 1,273 data were used for hyperparameter tuning. We used the MultilabelStratifiedKFold function from the iterstrat library for multi-label classification. The entire dataset was divided into five folds, ensuring all data were included in the validation set at least once. The average performance across these folds was reported as the final result.

Duration and Age Extraction Prompt

I will show you a text in which an anonymous user describes their mental status. Based on this, add labels according to the following criteria:

Duration: Does the given text mention a period of time when the user experienced psychiatric symptoms?
2: The user has been experiencing symptoms for a month or more. This is typically expressed in the text as "for a year","for over a month","for a long time" etc.
1: The user has been experiencing symptoms for less than a month. This is typically expressed in the text as "in recent days","recently","a week" etc.
0: The duration is not mentioned in the given text.

Age: Does the text mention the user's age or age range?
2: The user is in their 20s or older, which means the user is an adult and is referred to in the text as "adult","college student", "housewife" etc.
1: The user is a teenager or younger, which means the user is a child or young adult and is mentioned in the text as "middle school student", "high school student", " sophomore"or a specific age.
0: There is no mention of age in the text.

Let's think step by step.
1. Identify the symptoms the user is experiencing.
2. Determine how long the user has experienced these symptoms.
3. Determine user's age.

If you output 1 or 2, you must provide clear evidence. Output 0 if there are no mentions, and "No mention" for the reason.

EXAMPLE
Post: "...."
Answer: "...."

Figure 6: Prompt template of Duration and Age

C.2 Hyperparameter Settings

The experiments are conducted on two NVIDIA Quadro RTX A5000 GPUs, each with 24 GB of memory. The specific hyperparameter for all models are summarised below.

- **SVM+TF-IDF** used a minimum document frequency (min_df) of 2 and a maximum of 6000 features.
- **PsyEx** used the default settings following existing implementations, with detailed parameters similar to BERT-based models.
- **BERT-based models** used the $AdamW$ optimizer, a learning rate of 3e-05, klue/bert-base(pre-trained on the Korean corpus), a max length of 512, an early stop step of 3 epochs, and a batch size of 32.
- **Symp-based models** used the $AdamW$ optimizer, a learning rate of 0.01, Our data has 1 post per user, so we set the filter size of 1, filter number of 64, dropout rate of 0.2, an early stop step of 5 epochs, and a batch size of 64.

Affect Extraction Prompt

I will show you a text in which an anonymous user describes their mental status. Based on this, add labels according to the following criteria:

1. **Symptom**: What psychiatric symptoms or issues is the user experiencing?
2. **Social affect**: Does the symptom the user is experiencing interfere with the user's social functioning?
3. **Educational affect**: Does the symptom the user is experiencing interfere with the user's educational functioning?
4. **Occupational affect**: Does the symptom the user is experiencing interfere with the user's occupational functioning?
5. **Life-threatening affect**: Are the user's symptoms life-threatening?

For each affect:
1: The user's symptoms are currently affecting their social, educational, or occupational functioning, or are life-threatening.
0: There is no mention of educational, occupational, social, or life-threatening effects in the given text.

Let's think about it in steps.
1. Identify the symptoms the user is experiencing.
2. Determine how those symptoms have consequently affected their interpersonal, work, or academic life, or if they are life-threatening.

If you output 1, you must provide clear evidence from the text. Output 0 if there are no mentions, and "No mention" for the reason.

EXAMPLE
Post: "...."
Answer: "...."

Figure 7: Prompt template of Affect

- **LLM-based models** used the temperature 0 for GPT-3.5 and GPT-4o, using the same setting for factor extraction. MentalLLaMA used a temperature of 0.01.
- **CURE** used the $AdamW$ optimizer, a learning rate of 0.003, a batch size of 64, an MLP with a hidden dimension of 64, and the activation function elu (Clevert et al., 2015).

C.3 Template for Mental Disorder Detection

In Section 5.5.2, we compared LLM-based models for detecting mental disorders in a few-shot setting. We used three large language models: GPT-3.5, GPT-4o, and MentalLLaMA. When using MentalLLaMa, we translated user-generated posts into English because MentalLLaMa only works for English-written text. For GPT-based models that support multiple languages, we used the original Korean-written posts. The same prompt was applied to all models, as shown in Figure 8.

D Results on Symptom Identification

D.1 Performance on Symptom Identification

We developed a symptom identification model using BERT to detect psychiatric symptoms in user posts. For a comparison, two BERT-based models that were optimized for the Korean corpus were employed. Utilizing the representational capabilities of BERT, the model predicts the presence of 28 symptoms in user-generated content. The average scores across all symptoms are presented in Table 11.

LLM-based models Prompt

You're a mental health professional.
Based on the questions given, infer the disorder the user is currently experiencing. Your inference should be based on DSM-5 diagnostic criteria.

The inference results are displayed separately for each of the five categories. The five categories include four mental disease (depression disorders, anxiety disorders, sleep disorders, and eating disorders) and Non-Disease.

Non-Disease can include cases where the user appears to be experiencing a condition outside of the four target conditions, or where it is unknown what condition the user is experiencing, or cases where the user has psychiatric symptoms but is not considered a mental disease.

Here are some examples.

EXAMPLE 1
Post: "...."
Answer: "...."

EXAMPLE 2
Post: "...."
Answer: "...."

EXAMPLE 3
Post: "...."
Answer: "...."

Figure 8: Prompt of Mental Disorder Detection for LLM-based models.

Model	Prec.	Rec.	F1.
BERT	0.854	0.780	0.809
RoBERTa	0.859	0.765	0.800

Table 11: Symptom identification results on KoMOS. The average performance scores across all symptoms are reported.

D.2 Detailed Results for Each Symptoms

We reported the detection results for each of the 28 symptoms in Table 12. The results from the BERT model, which demonstrated better performance than other BERT-based models, are presented. There is a performance imbalance in each symptom class, which may be due to the differences in the number of instances in each class or the different complexity of the symptom patterns expressed in user-generated posts, which makes symptom identification challenging.

E Results for Mental Disorder Detection

E.1 Ablation Study of Context Factor

We evaluated whether our proposed context factors can help in disease detection. We measured the performance across five mental disorder categories by removing each contextual factor one by one from the Symp$_{context}$ model (which considers both

Symptom	F1-score	Support
Palpitations	0.95	155
Chest discomfort	0.92	234
Hypersomnia	0.77	64
Irritability	0.76	209
Tremor	0.91	85
Sweating	0.96	64
Abdominal discomfort	0.86	80
Insomnia	0.87	226
Feeling anxious	0.93	467
Other symptoms due to anxiety	0.83	303
Social anxiety	0.78	75
Sensory disturbances during sleep	0.76	39
Parasomnias	0.85	36
Sleep paralysis	0.74	23
Poor quality of sleep	0.77	118
Loss of appetite	0.79	64
Dietary restrictions	0.41	46
Paresthesia	0.64	83
Depressed mood	0.91	345
Difficulty in eating	0.70	46
Suicidal ideation	0.89	155
Self-harm	0.85	36
Self-induced vomiting	0.90	97
Excessive sleepiness	0.64	34
Fear of weight gain	0.83	104
Phobias	0.69	84
Binge eating episodes	0.91	130
Loss of energy	0.85	188

Table 12: Performance of symptom identification for each 28 symptoms.

symptom and context categories). The detailed performance of the factor ablation study is presented in Table 14. As shown in Table, each context factor plays a different role in detecting different diseases. For example, the exclusion of the frequency leads to a performance decrease in most categories (e.g., Anxiety Disorder: $0.891 \rightarrow 0.876$ and Sleep Disorder $0.863 \rightarrow 0.846$, in terms of recall). On the other hand, in the case of Depressive Disorder, performance improved when the frequency was removed ($0.812 \rightarrow 0.830$ on recall). These findings suggest that context factors contribute to improve the performance of the disorder detection with different roles.

E.2 Performance on LLM-based Methods

We assessed the performance of LLM-based models across five categories of mental disorders. The detailed performance of the LLM-based model is presented in Table 14. As shown Table 14, large language models tend to over-diagnose diseases, resulting in high recall values for disease cases but lower performance in non-disease cases. Additionally, the enhanced capability of the large language models do not significantly impact the performance difference. Despite the fact that GPT-4o is an upgraded version of GPT-3.5, the performance gap between the two models is marginal, except for non-disease cases. In non-disease cases, GPT-4o demonstrated better predictive performance than GPT-3.5, indicating that more advanced language models can accurately detect distinctions between disease and non-disease classes based on expert knowledge and language capabilities.

E.3 Statistical Analysis for Performance

we conducted the paired t-test to verify the statistical significance of the experiment result. We report the results of 5-fold cross-validation, based on the differences in average recall and average F1 score between our model and the baselines. We verified that all the p-values of all tests are below 0.05, indicating that the results are statistically significant. The result is shown in Table 13

Model	Recall			F1-score		
	Average	T-statistic	P-value	Average	T-statistic	P-value
BERT	0.799±0.010	-5.72	0.0046	0.801±0.007	-6.71	0.0026
Symp	0.790±0.006	-6.89	0.0023	0.792±0.006	-8.86	0.0009
PsyEx	0.785±0.016	-4.45	0.0113	0.794±0.009	-4.78	0.0088
GPT4o	0.843±0.002	4.92	0.0080	0.702±0.010	-14.17	0.0001

Table 13: Results for paired T-Test between baselines and the proposed method.

F Error Analysis

We report two representative error cases with description of when these cases are made and how the errors are propagated.

F.1 Strong agreement among sub-models to incorrect label

The most errors of the proposed framework were made when all the sub-models show a strong agreement for the incorrect disease. In Table 15, the doctors who labeled this case explained that ordinary people may also have the same feelings in the described situation, so the symptoms are not severe enough to be diagnosed as a phobia. From the model perspective, Anxiety Disorder was a final decision since the decisions of all the sub-models are the same as Anxiety Disorder. The reason for the incorrect decision by each sub-model is due to the failure of capturing the information of specific symptoms for the given disease and their severity. In this example, the sub-models fail to acknowledge that sweating is not an extreme symptom for anxiety disorder. We expect that this error can be

Model	Depressive Disorders			Anxiety Disorders			Sleep Disorders			Eating Disorders			Non-disease		
	Pre.	Rec.	F1.	Pre.	Rec.	F1.	Pre.	Rec.	F1.	Pre.	Rec.	F1.	Pre.	Rec.	F1.
Symp$_{context}$	**0.802**	0.812	0.807	0.803	**0.891**	**0.845**	0.798	**0.863**	**0.829**	0.878	**0.933**	0.905	0.778	**0.484**	**0.595**
w/o duration	0.790	0.825	0.807	0.804	0.875	0.838	0.808	0.846	0.826	0.879	0.924	0.901	0.780	0.479	0.593
w/o age	0.794	0.826	**0.809**	**0.812**	0.868	0.839	0.814	0.832	0.822	**0.892**	0.917	0.904	0.763	0.479	0.588
w/o cause	0.800	0.811	0.805	0.809	0.873	0.840	0.806	0.840	0.822	0.891	0.931	0.910	**0.787**	0.465	0.583
w/o frequency	0.788	**0.830**	0.808	0.809	0.876	0.841	0.812	0.846	0.828	0.883	0.915	0.899	**0.787**	0.463	0.582
w/o affect	0.798	0.820	**0.809**	0.809	0.872	0.839	**0.815**	0.827	0.821	**0.892**	**0.933**	**0.912**	0.781	0.465	0.582
GPT-3.5	**0.682**	0.907	**0.778**	0.553	0.972	0.704	**0.705**	0.790	**0.745**	**0.753**	**0.995**	**0.857**	0.717	0.235	0.353
GPT-4o	0.614	**0.940**	0.743	**0.572**	**0.990**	**0.725**	0.585	**0.984**	0.733	0.752	0.993	0.856	**0.840**	0.310	**0.453**
MentalLLaMa	0.247	0.898	0.387	0.434	0.494	0.462	0.213	0.902	0.344	0.178	0.699	0.284	0.290	**0.453**	0.353

Table 14: Detailed results for mental disorder detection. The first section represents results of ablation study for context factor, while the second section represents performance of large language models.

Post	BERT$_{question}$	BERT$_{context}$	Symp$_{symptom}$	Symp$_{context}$	CURE$_{ours}$	Ground Truth
I think I might have a phobia. When I stand on glass floors, I get thoughts that the glass might break. This makes me sweat and feel anxious. Why am I like this?	Anxiety Disorder - A: 0.95	Anxiety Disorder - A: 0.91	Anxiety Disorder - A: 0.95	Anxiety Disorder - A: 0.94	Anxiety Disorder - A : 0.99	Non-Disease (-)
I'm so tired that I'm dozing off while working. Sometimes, I confuse the work I've done while drowsy with things that happened in a dream. I'm wondering if this might be a problem serious enough to warrant a visit to the doctor.	Non-Disease - S: 0.18 - N: 0.83	Sleep Disorder - S: 0.63 - N: 0.41	Sleep Disorder - S: 0.62 - N: 0.38	Sleep Disorder - S: 0.62 - N: 0.47	Non-Disease - S: 0.15 - N: 0.96	Sleep Disorder (-)

Table 15: An analysis of the error cases for the proposed method. Two cases are considered for representing error propagation within the proposed method. The value under the predicted label for each model denotes the predictive logits corresponding to the disease class.

addressed by designing more sophisticated context factors such as types of symptoms (of each disease) with their severity.

F.2 Following the decision by sub-models with high confidence

The second-most error case is that the final decision is made as the output of the sub-model with high confidence when the sub-models made different predictions with different confidence. An example case is described in the Table 15. Although the majority of sub-models outputs Sleep Disorder, the final decision by the proposed model is Non-Disease. The pair of numerical values in the parentheses are the predictive logits of each class, which can be used to estimate the model's confidence in the decision of each label. In this case, three models predicting Sleep Disorder show lower confidence, whose values are in [0.4, 0.6]. In contrast, the confidence of the BERT$_{question}$ is much higher than the others (more than 0.8). The final output of the proposed model is Non-Disease, which seems to be taken from the decision by BERT$_{question}$.