Process Knowledge Infused Learning for Clinician Friendly Explanations

Kaushik Roy[1], **Yuxin Zi**[1], **Manas Gaur**[2] , **Jinendra Malekar**[1], **Qi Zhang**[1], **Vignesh Narayanan**[1], **Amit Sheth**[1]

[1] Artificial Intelligence Institute, University of South Carolina
Columbia, South Carolina, US
[2] University of Maryland, Baltimore County, US
{kaushikr, yzi}@email.sc.edu, manas@umbc.edu, jmalekar@mailbox.sc.edu, qz5@cse.sc.edu,{vignar,amit}@sc.edu

Abstract

Language models have the potential to assess mental health using social media data. By analyzing online posts and conversations, these models can detect patterns indicating mental health conditions like depression, anxiety, or suicidal thoughts. They examine keywords, language markers, and sentiment to gain insights into an individual's mental well-being. This information is crucial for early detection, intervention, and support, improving mental health care and prevention strategies. However, using language models for mental health assessments from social media has two limitations: (1) They do not compare posts against clinicians' diagnostic processes, and (2) It's challenging to explain language model outputs using concepts that the clinician can understand, i.e., clinician-friendly explanations. In this study, we introduce Process Knowledge-infused Learning (PK-iL), a new learning paradigm that layers clinical process knowledge structures on language model outputs, enabling clinician-friendly explanations of the underlying language model predictions. We rigorously test our methods on existing benchmark datasets, augmented with such clinical process knowledge, and release a new dataset for assessing suicidality. PK-iL performs competitively, achieving a 70% agreement with users, while other XAI methods only achieve 47% agreement (average inter-rater agreement of 0.72). Our evaluations demonstrate that PK-iL effectively explains model predictions to clinicians.

Introduction

A long-standing problem in adopting language models for clinician assistance has been the lack of clinician-friendly explanations for the model's predictions [1]. In practice, a clinical guideline or process is often detailed by which the clinician can assess or label patients. For example, to label patients for degrees of suicidal tendencies in a physical clinical setting, a well-known scale, the Columbia Suicide Severity Rating Scale (CSSRS) (Bjureberg et al. 2021), is used to determine the right set of labels. The green part of Figure 1 (b) shows the CSSRS scale, a *process*, which consists of six conditions whose values determine four assessment outcomes from the set {*indication*, *ideation*, *behavior*, *attempt*}. Similarly, when patients are assessed for depression, clinicians evaluate patient responses against a process or guideline like the Patient Health Questionnaire-9 (PHQ-9) and provide explanations for their assessment using the same. The blue part of Figure 1 (b) shows the PHQ-9 assessment process. Language models do not explicitly leverage such process knowledge to derive their predictions. Furthermore, language model predictions are typically explained using XAI methods, such as LIME and SHAP, which fits a simpler interpretable surrogate model (Adadi and Berrada 2018; Ribeiro, Singh, and Guestrin 2016; Lundberg and Lee 2017). XAI models, however, provide explanations that benefit computer scientists in debugging and improving language models but are of limited utility to the clinician for making decisions. Additionally, it is challenging to approximate very large and complex models, e.g., language models (LMs) using simpler surrogate models (Vaswani et al. 2017).

We propose a novel learning framework *Process Knowledge infused Learning* (PKiL) that leverages explicit representations of publicly available knowledge of processes and guidelines to augment language models to enable clinician-friendly explanations. Crucially, PKiL incorporates process knowledge structures to provide explanations for model predictions using concepts that are familiar to a clincian. Figure 1 shows the execution flow of a model trained using PKiL. The PKiL learning framework achieves this through a novel training method with the following salient features - (1) PKiL leverages powerful language models with hundreds of millions of parameters while requiring training of very few additional parameters (equal to the number of process knowledge conditions, e.g., conditions in Figure 1 (b)) to obtain clinician-friendly explanations, (2) The optimization objective is simple to understand, enabling globally optimal solution discovery through various optimization procedures.

Problem Formulation, Resource Construction, and Process Knowledge Infused Learning

Problem Formulation

Let $X_{\mathcal{D}}$ denote a dataset of input texts and their labels in a domain $\mathcal{D}$. An example of an input post is shown in Figure 1 (a), and its suicidality assessment label is from the set {*indication*, *ideation*, *behavior*, *attempt*} in the domain of mental health. Let $Pk_{\mathcal{D}}$ denote the relevant process knowledge available to us from established literature in domain

[1]https://globelynews.com/world/chatgpt-ai-ethics-healthcare/

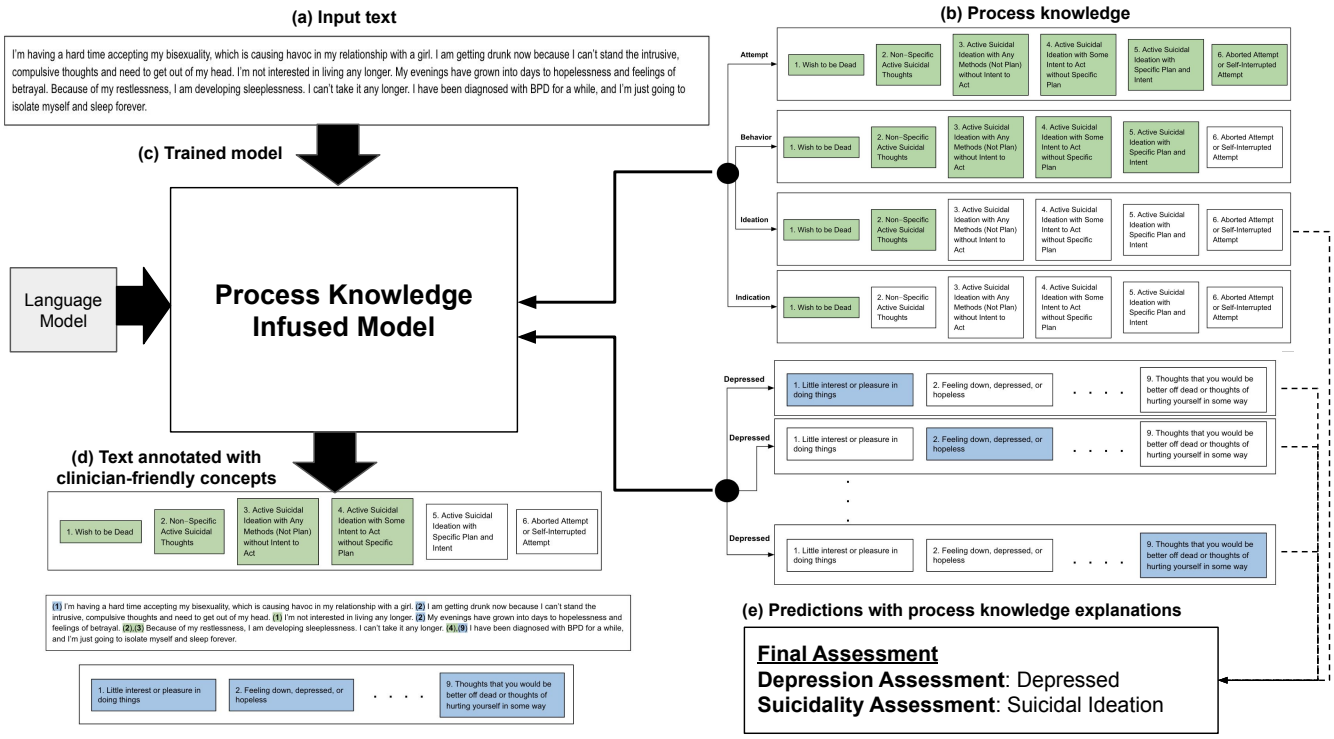

Figure 1: Overview of PKiL inference for an input text. The model uses two arguments, the input text (a) and the process knowledge (b). The process knowledge shows conditions that must be satisfied for a given label. The green part shows process knowledge conditions for suicidality assessment, and the blue part shows the same for depression assessment. For example, for the label *attempt* in the suicidality assessment process knowledge, all conditions 1-6 need to be satisfied. For the label *indication*, only condition 1 needs to be satisfied. The model then annotates text fragments with clinician-friendly concepts from the process knowledge, as shown in (d). The final assessment predictions are obtained through the relevant process knowledge conditions that apply, as shown in (e).

$\mathcal{D}$. For example, Figure 1(b) shows the process of obtaining suicidality assessment labels. Let $\Lambda_{\mathcal{D}}$ be a language model available to us that is fine-tuned on domain specific data (e.g., BERT fine-tuned on mental health posts from social media). Process Knowledge infused Learning (PKiL) is a training method that makes combined use of $X_{\mathcal{D}}$ and $Pk_{\mathcal{D}}$ to evaluate the conditions in the process knowledge to predict the final label. The evaluated conditions in the process knowledge are familiar to clinicians and therefore enable clinician-friendly explanations for predictions, as shown in Figure 1.

Resource Construction - Construction of Process Knowledge Augmented Datasets

Due to the recent push for explainable and trustworthy AI, recent studies have published new datasets with knowledge of established processes and guidelines used in a particular domain. For example, Gupta et al. constructed the PRIMATE dataset, which includes a series of depression-related posts labeled by human annotators by checking against the PHQ-9 depression assessment process knowledge (Gupta et al. 2022). Roy et al. construct the ProKnow dataset that consists of similar process knowledge for question generation (e.g., generate questions about symptoms before causes) while eliciting mental health-related conversation for psychometric test evaluations (Roy et al. 2023). We call such datasets process knowledge augmented datasets (Sheth et al. 2022). Gaur et al. used the CSSRS, the suicidality assessment process knowledge, to annotate suicidality labels for a set of Reddit posts extracted from suicidality-related subreddits (Gaur et al. 2021). We will call this dataset CSSRS 1.0, an example of $X_{\mathcal{D}}$ in our problem formulation. Even though their work labeled the posts using the process knowledge contained in the CSSRS as annotation guidelines, the exact process knowledge $Pk_{\mathcal{D}}$ used per data point was not stored. Therefore, we first obtain the $Pk_{\mathcal{D}}$ using the following procedure:

1. First, we fine-tune the models Word2Vec, SBERT, RoBERTa, T5, ERNIE, and Longformer on the CSSRS 1.0 dataset, i.e., the $\Lambda_{\mathcal{D}}$ in our formulation (Mikolov et al. 2013; Reimers and Gurevych 2019; Liu et al. 2019; Raffel et al. 2020; Zhang et al. 2019; Beltagy, Peters, and Cohan 2020).
2. Second, we evaluate each post in the CSSRS 1.0 dataset

against the CSSRS $Pk_{\mathcal{D}}$ conditions using cosine similarity between the fine-tuned representations of the posts and the conditions. Condition evaluation returns 1.0 if the condition is satisfied, else 0.0. We set the similarity threshold to 0.5. We do this for all the models and use the max similarity that is greater than the threshold of 0.5.

3. Next, we obtain a label for each post in $X_{\mathcal{D}}$ from the set {*indication*, *ideation*, *behavior*, *attempt*} by comparing the evaluated condition values against the $Pk_{\mathcal{D}}$. For example, if only condition 1, which is *wish to be dead* evaluates to 1.0, the label is *indication* (see Figure 1 (b)).
4. Lastly, we provide our labels to three domain experts and task them with either retaining the labels or editing the labels by referring to the CSSRS $Pk_{\mathcal{D}}$ while recording the inter-rater agreement.

The domain experts in the study checked through the labels of 448 Reddit posts in $X_{\mathcal{D}}$. They edited 235/448 posts and provided the relevant process knowledge conditions 1-6, evaluated during the edit. A substantial inter-rater agreement of 0.84 was recorded. Crucially, we augment the CSSRS 1.0 to include the specific process knowledge used for the edited label. We call this new dataset CSSRS 2.0. Examples from the dataset can be found at the link in the footnote[2]. We will use $X_{\mathcal{D}}^{Pk}$ to denote process knowledge augmented datasets. Note that $|X_{\mathcal{D}}^{Pk}| \leq |X_{\mathcal{D}}|$. For example, CSSRS 2.0 has 235 data points, whereas CSSRS 1.0 has 448 data points. Our experiments use CSSRS 2.0 and PRIMATE.

Process Knowledge Infused Learning

Consider a single condition process knowledge $Pk_{\mathcal{D}}$ to predict a binary label L for an input $x \in X_{\mathcal{D}}^{Pk}$:

$$\begin{aligned} if\ (C(x) = 1), L(x) = 1 \\ else, L(x) = 0 \end{aligned}$$

Here $C(x)$ is a condition evaluation function for the input x that evaluates to 1.0 if the condition is satisfied and 0 if the condition is not satisfied. $Pk_{\mathcal{D}}$ can be written algebraically as:

$$\begin{aligned} L(x) = \mathbf{I}(L(x) = 1)(C(x) = 1) \\ + \mathbf{I}(L(x) = 0) \end{aligned} \tag{1}$$

Here $\mathbf{I}(L(x) = l)$ is the indicator function that evaluates to 1 or 0, indicating whether the value that the label $L(x)$ takes is equal to l. How do we mathematically formulate $C(x) = 1$? We can parameterize $C(x) = 1$ as $S(e_x^{\Lambda_{\mathcal{D}}}, e_C^{\Lambda_{\mathcal{D}}}) \geq \theta_C$, where S is a similarity function (e.g., cosine similarity) and θ_C is the similarity threshold. The $e_x^{\Lambda_{\mathcal{D}}}$ and $e_C^{\Lambda_{\mathcal{D}}}$ are embeddings of the input and condition obtained using a domain-specific fine-tuned language model $\Lambda_{\mathcal{D}}$. Thus, we can write a parameterized approximation to (1) as:

$$\begin{aligned} \hat{L}(x, \theta_C) = \mathbf{I}(L(x) = 1) S(e_x^{\Lambda_{\mathcal{D}}}, e_C^{\Lambda_{\mathcal{D}}}) \geq \theta_C) \\ + \mathbf{I}(L(x) = 0) \end{aligned} \tag{2}$$

Now we consider a slightly more complex process knowledge $Pk_{\mathcal{D}}$, a multilabel and multi-conditioned process knowledge to predict label $L \in \{1, 2, 3\}$, given conditions $C1, C2, C3$, for an input $x \in X_{\mathcal{D}}^{Pk}$:

$$\begin{aligned} if\ (C1(x) = 1 \wedge C2(x) = 1), L(x) = 1 \\ if\ (C1(x) = 1 \wedge C3(x) = 1), L(x) = 2 \\ else, L(x) = 3 \end{aligned}$$

Similar to (1), we can write this $Pk_{\mathcal{D}}$ algebraically as:

$$\begin{aligned} L(x) = \mathbf{I}(L(x) = 1)(C1(x) = 1)(C2(x) = 1) \\ + \mathbf{I}(L(x) = 2)(C1(x) = 1)(C3(x) = 1) \\ + \mathbf{I}(L(x) = 3) \end{aligned} \tag{3}$$

Following a similar procedure as the one used to derive (2), we obtain:

$$\begin{aligned} &\hat{L}(x, \theta_{C1}, \theta_{C2}) = \\ &\mathbf{I}(L(x) = 1)(S(e_x^{\Lambda_{\mathcal{D}}}, e_{C1}^{\Lambda_{\mathcal{D}}}) \geq \theta_{C1})) \\ &(S(e_x^{\Lambda_{\mathcal{D}}}, e_{C2}^{\Lambda_{\mathcal{D}}}) \geq \theta_{C2})) \\ &+ \mathbf{I}(L(x) = 2)(S(e_x^{\Lambda_{\mathcal{D}}}, e_{C1}^{\Lambda_{\mathcal{D}}}) \geq \theta_{C1}) \\ &(S(e_x^{\Lambda_{\mathcal{D}}}, e_{C3}^{\Lambda_{\mathcal{D}}}) \geq \theta_{C3})) \\ &+ \mathbf{I}(L(x) = 3) \end{aligned} \tag{4}$$

Generally, given multi-condition process knowledge $Pk_{\mathcal{D}}$ for multilabel prediction of the form

$$if\ \wedge_j\ (C_j(x) = 1), L(x) = l$$

we get its algebraic form as

$$L(x) = \mathbf{I}(L(x) = l) \prod_j (C_j(x) = 1) \tag{5}$$

Denoting all the parameters as the set $\{\theta_{C_j}\}$ we get the parameterization

$$\hat{L}(x, \{\theta_{C_j}\}) = \mathbf{I}(L(x) = l) \prod_j (S(e_x^{\Lambda_{\mathcal{D}}}, e_{C_j}^{\Lambda_{\mathcal{D}}}) \geq \theta_{C_j}) \tag{6}$$

For all $x \in X_{\mathcal{D}}^{Pk}$, we get a system of equations like (6).

Sentiment Analysis The conditions in the process knowledge help the model assess problem issues. However, a complete mental health assessment usually also involves the identification of signs of positivity. Therefore for each θ_{C_j}, we also optimize for a γ_{C_j} term, where the model predicts positive sentiment in the input if $S(e_x^{\Lambda_{\mathcal{D}}}, e_{C_j}^{\Lambda_{\mathcal{D}}}) \leq \theta_{C_j} + \gamma_{C_j}$.

Optimization Problem Formulation For a process knowledge augmented dataset $X_{\mathcal{D}}^{Pk}$, we know the ground truths $L(x)$ for all $x \in X_{\mathcal{D}}^{Pk}$. We want to solve for the unknown parameters θ_{C_j} that yields minimum error between the parameterized approximation $L(x, \{\theta_{C_j}\})$ and the ground truth $L(x)$ i.e.,

$$\sum_{x \in X_{\mathcal{D}}^{Pk}} \mathcal{E}(\hat{L}(x, \{\theta_{C_j}\}), L(x))$$

Here $\mathcal{E}$ denotes the error function. The choice of similarity functions S is a hyperparameter (We explore cosine similarity and normalized Gaussian kernels in our experiments).

[2]https://anonymous.4open.science/r/MenatalHealthAnoynomous-8CC3/cssrs\%202.0.csv

Projected Newton's method: When one of the $\{\theta_{C_j}\}$ are fixed, setting $\mathcal{E}(\hat{L}(x, \{\theta_{C_j}\}), L(x))$ to be the cross entropy loss reduces to a strongly convex objective that can be solved by **Newton's method** (with ε corrections for low determinant Hessians). After each optimization step, we project the θ_{C_j} to the $[-1, 1]$ range.

Grid Search: Since the number of parameters to optimize is small (six for CSSRS 2.0 and nine for PRIMATE), we can perform a grid search over a predefined set of grid values to find the values that yield minimum cross entropy loss. For our choice of S, we choose **cosine similarity and normalized Gaussian kernel**; therefore, grid search candidate values are in the $[-1, 1]$ range.

Optimizing for the γ_{C_j}: To find the optimal γ_{C_j}, we first predict positive and negative sentiment labels using the **Stanford CoreNLP** model for all the inputs. Next, we perform a grid search in the $[-1, 1]$ range and set values for the γ_{C_j} that results in the maximum agreement between $S(e_x^{\Lambda_\mathcal{D}}, e_{C_j}^{\Lambda_\mathcal{D}}) \leq \theta_{C_j} + \gamma_{C_j}$ and the Stanford CoreNLP model labels (only the positive labels).

In our experiments, we try both Newton's method and grid search optimization strategies.

Experiments and Results

We demonstrate the effectiveness of PkiL training using PRIMATE and CSSRS 2.0 combined with several state-of-the-art language models. We also perform experiments with prompting Text-Davinci-003 using the langchain library[3].

Process Knowledge Augmented Datasets

For CSSRS 2.0, the process knowledge is shown in Figure 1 (b) (the green part). We input this process knowledge in the form[4]:

$$
\begin{aligned}
&if\ ((C1(x), C2(x), C3(x), C4(x), C5(x), C6(x)) = 1),\\
&\ L(x) = attempt\\
&if\ ((C1(x), C2(x), C3(x), C4(x), C5(x)) = 1),\\
&\ L(x) = behavior\\
&if\ ((C1(x), C2(x)) = 1), L(x) = ideation\\
&if\ (C1(x) = 1), L(x) = indication
\end{aligned}
$$

The conditions $C1 - C6$ in the CSSRS are:

$C1:\ Wish\ to\ be\ dead$

$C2:\ Non - Specific\ Active\ Suicidal\ Thoughts$

$C3:\ Active\ Suicidal\ Ideation\ with\ Any\ Methods$
$(Not\ Plan)\ without\ Intent\ to\ Act$

$C4:\ Active\ Suicidal\ Ideation\ with\ Some\ Intent\ to\ Act$
$without\ Specific\ Plan$

$C5:\ ActiveSuicidal\ Ideation\ with\ Specific\ Plan\ and\ Intent$

$C6:\ Aborted\ Attempt\ or\ Self - Interrupted\ Attempt$

For PRIMATE, the process knowledge is a set of nine conditions. If any of the conditions evaluate to yes, the depression assessment label is 1. This is a binary classification task. We input this process knowledge in the form (we collapse conditions $C3 - C8$ for brevity):

$$
\begin{aligned}
&if\ (C1(x) = 1), L(x) = 1\\
&if\ (C2(x) = 1), L(x) = 1\\
&\ldots\\
&if\ (C9(x) = 1), L(x) = 1\\
&else, L(x) = 0
\end{aligned}
$$

The conditions $C1 - C9$ in the PHQ-9 are:

$C1:\ Little\ interest\ or\ pleasure\ in\ doing\ things$

$C2:\ Feelingdown,\ depressed,\ or\ hopeless$

$C3:\ Trouble\ falling\ or\ staying\ asleep,$
$or\ sleeping\ too\ much$

$C4:\ Feeling\ tired\ or\ having\ little\ energy$

$C5:\ Poor\ appetite\ or\ overeating$

$C6:\ Feeling\ bad\ about\ yourself,$
$or\ that\ you\ are\ a\ failure,$
$or\ have\ let\ yourself\ or\ your\ family\ down$

$C7:\ Trouble\ concentrating\ on\ things,$
$such\ as\ reading\ the\ newspaper\ or\ watching\ television$

$C8:\ Moving\ or\ speaking\ so\ slowly\ that$
$other\ people\ could\ have\ noticed$
$Or\ so\ fidgety\ or\ restless\ that$
$you\ have\ been\ moving\ a\ lot\ more\ than\ usual$

$C9:\ Thoughts\ that\ you\ would\ be\ better\ off\ dead$
$or\ thoughts\ of\ hurting\ yourself\ in\ some\ way?$

Examples from the PRIMATE dataset can be found at the link in the footnote [5].

Experimental and Hyperparameter Configurations During Training

1. **Embedding models for input post and questions:** We use the models Word2Vec, SBERT, RoBERTa, T5, ERNIE, and Longformer fine-tuned on the training data.
2. **Similarity function:** We explore the cosine similarity and the normalized Gaussian kernel (input vectors are normalized to be unit length before plugging into the Gaussian kernel). For the normalized Gaussian kernel, we range the scale parameter between $[-1, 1]$ in increments of 0.001.
3. **Parameters for grid search:** During grid search optimization, we explore parameters in the $[-1, 1]$ range, again in increments of 0.001.
4. **No. of epochs for Newton's optimization method:** We set max epochs of 100 and experiment with batch sizes of 16 and 32 for Newton's method. We train for only 100 epochs as we have far more equations than unknowns and also perform early stopping if the total parameter differences are less than 0.001.

[3] https://langchain.readthedocs.io/en/latest/

[4] Examples can be found at the link: https://anonymous.4open.science/r/MenatalHealthAnoynomous-8CC3/cssrs_annotate.txt

[5] https://github.com/primate-mh/Primate2022

Text-Davinci-003 Experiment Details

We use the langchain library and write a prompt template to obtain answers to the process knowledge questions from Text-Davinci-003. For example, Figure 2 shows the prompt template for the first condition $C1: \ Wish\ to\ be\ dead$ from the CSSRS process knowledge. For sentiment analysis, we set the *question* variable in Figure 2 to *positive sentiment*. We will call this model Text-Davinci-003$_{\text{PK}}$. Once we evaluate all the conditions, we follow the process knowledge pertaining to the evaluated condition values to determine the label.

```
from langchain.prompts import PromptTemplate
input = <INPUT TEXT>
symptom = "Wish to be dead"
template = "Is the input post {input}, exhibiting {symptom}"
prompt = PromptTemplate(input_variables = ["input","symptom"],
                        template = template)
```

Figure 2: Using the langchain library to prompt Text-Davinci-003 for answers to questions from the process knowledge.

Quantitative Results and Discussion

Figure 3 shows the results of PKiL for various experiment configurations for the CSSRS 2.0 and PRIMATE datasets. The figure also shows results from the Text-Davinci-003$_{\text{PK}}$ model.

Quantitative Results for CSSRS 2.0: First, excluding the Text-Davinci-003$_{\text{PK}}$ from the analyses, we observe that SBERT trained using PKiL with a normalized Gaussian kernel performs the best in terms of accuracy, and the Word2Vec model performs the best on AUC-ROC scores for the CSSRS 2.0 dataset. In general, we see that PKiL leads to large boosts in performance of up to 14% over the baseline. Analysis of The Text-Davinci-003$_{\text{PK}}$ model performance reveals that it is the best performer among all the models for the CSSRS 2.0 dataset. Our experiments show that large language models can significantly increase suicidality assessment performance when leveraging process knowledge structures and process knowledge-augmented datasets.

Quantitative results for PRIMATE: Again, first excluding the Text-Davinci-003$_{\text{PK}}$ from the analyses, we observe that RoBERTa trained using PKiL with a cosine similarity function performs the best in terms of accuracy, and SBERT and ERNIE perform the best on AUC-ROC scores for the PRIMATE dataset. In general, we see that PKiL leads to large boosts in performance of up to 23% over the baseline. Analysis of The Text-Davinci-003$_{\text{PK}}$ model performance reveals that it is the best performer in terms of accuracy among all the models for the PRIMATE dataset. Our experiments show that large language models can also significantly increase depression assessment performance when leveraging process knowledge structures and process knowledge augmented datasets.

Qualitative Results and Discussion

We evaluate PkiL model outputs qualitatively for the following aspects:

- **Mental health disturbance assessment**: The final label predicted by the model, i.e., the label *depression* for depression assessment), and a label from the set {*indication*, *ideation*, *behavior*, *attempt*} for suicidality assessment.
- **PHQ-9 depression concepts identified**: A list of concepts resulting from evaluating conditions C1-C9 using the learned thresholds θ_{C_j}. For the Text-Davinci-003$_{\text{PK}}$ model, we prompt the model using code as shown in Figure 2.
- **CSSRS suicidality concepts identified**: A list of concepts resulting from evaluating conditions C1-C6 using the learned thresholds θ_{C_j}. Similar to the depression case, for the Text-Davinci-003$_{\text{PK}}$ model, we prompt the model using code as shown in Figure 2.
- **Positive sentiment assessment**: Using the learned θ_j and γ_j to identify input post fragments that convey positive sentiment.

Baseline Model Explanations: We use the bert-viz visualization technique[6] to interpret the contributions of the different input post fragments to the prediction outcome (the `CLS` token). Figure 3(e) shows the output for SBERT. The highlights convey meaningful information from the perspective of depression, which is the correct label. However, it is unclear how the highlights map to clinician-friendly concepts from process knowledge guidelines for depression assessment. A manual post-processing layer for mapping to clinician-friendly concepts is needed in order to verify the prediction.

PKiL Model Explanations: We divide the input post into contiguous fragments of max size 3 sentences for models and infer the process knowledge condition values using the PKiL trained models and the parameters θ_{C_j} and θ_{γ_j}. We divide for enhanced clinician-friendly explainability as simply annotating the whole posts with concepts still requires additional post-processing by the human to glean out fragments that correspond to problem issues and positive sentiments. Figure 3(f) shows the output of the SBERT model trained using PKiL with the normalized Gaussian kernel. Figure 3(g) shows the output of prompting the Text-Davinci-003$_{\text{PK}}$ as shown in Figure 2. We can readily observe that the explanations are more useful to the clinician as they directly explain the outcome in terms of concepts used in everyday practice. Finally, we provided PKiL explanations to the experts who helped construct the CSSRS 2.0 dataset and asked them to provide the percentage of times they found the explanations beneficial. We also provided baseline explanations for comparison. In order to control for bias, we tell them that humans generate PKiL explanations, and language models generate the baseline explanations. PKiL explanations scored 70% vs 47% for the baseline models. We recorded an inter-annotator agreement of 0.72. We analyzed the 30% that the experts did not find beneficial and observed that models have difficulty distinguishing casual mentions from serious ones. For example, a Reddit user reported wanting to kill themselves out of class boredom before identifying a legitimate clinical issue much further into their post. We leave the investigation

[6] https://github.com/jessevig/bertviz

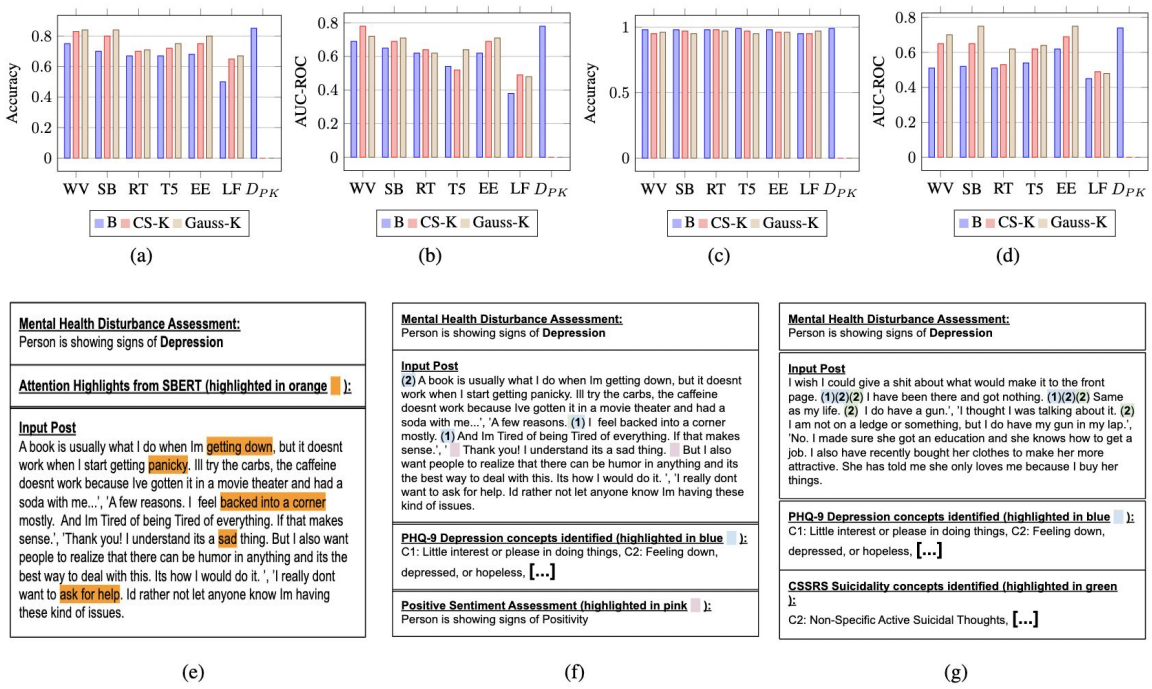

Figure 3: (a) and (b) present results for the CSSRS 2.0 dataset, while (c) and (d) show the results for the PRIMATE dataset. The mean accuracy/AUC-ROC of different language models (LMs) - Baseline fine-tuned model (B), PKiL performance with Cosine Similarity Kernel (CS-K), and PKiL performance with normalized Gaussian Kernel similarity (Gauss-K) - are displayed. The prompt-based model Text-Davinci-003$_{\text{PK}}$ model (D_{PK}) doesn't utilize CS-K or Gauss-K, so no associated bar is shown. WV: Word2Vec, SB: SBERT, RT: RoBERTa, EE: ERNIE, LF: LongFormer. (e) The self-attention-based interpretability visualization for the SBERT baseline model indicates correct predictions and sensible highlights. However, the mapping of these highlights to clinician-friendly concepts used in practice is unclear. Baseline language models consistently struggle to capture negation. (f) The SB model trained with PKiL using the normalized Gaussian kernel provides clinicians with annotated explanations that are more familiar. Additionally, the PKiL parameters enable the analysis of fragments conveying positive sentiment. (g) Explanations from the Text-Davinci-003$_{\text{PK}}$ model also demonstrate that leveraging process knowledge helps clinicians better understand the annotated explanations, as they are associated with familiar problem concepts.

of these posts for future work (e.g., by expanding our framework to detect sarcasm).

Conclusion

In this study, we develop a novel paradigm PKiL that leverages the combined benefits of explicit process knowledge and high-performance language models to provide predictions and explanations that the end user can readily understand. Our experiments demonstrate the effectiveness of PKiL both quantitatively and qualitatively. Such improved understanding of language model predictions can inform insights for refining existing process knowledge guidelines (e.g., adaptation to Reddit vocabulary) to facilitate remote monitoring and improved access to healthcare via social media platforms.

Reproducibility: We provide the trained model for SBERT with normalized Gaussian kernel similarity, the CSSRS 2.0 dataset, and the CSSRS process knowledge used in our experiments at the link in the footnote[7]. Additionally, we also provide a Python notebook for users to play with the Text-Davinci-003$_{\text{PK}}$ model at the link in this footnote[8].

Ethics Statement: We adhere to anonymity, data privacy, intended use, and practical implication of the AI-based mental health assessment systems. The clinical process knowledge does not contain personally identifiable information. The datasets covered in the survey are publicly available and can be obtained from user-author agreement forms. Figures and examples are abstract and do not represent real-time data sources or any person.

References

Adadi, A.; and Berrada, M. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI).

[7] https://anonymous.4open.science/r/MenatalHealthAnoynomous-8CC3/README.md

[8] https://anonymous.4open.science/r/MenatalHealthAnoynomous-8CC3/app.ipynb

IEEE access, 6: 52138–52160.

Beltagy, I.; Peters, M. E.; and Cohan, A. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Bjureberg, J.; Dahlin, M.; Carlborg, A.; Edberg, H.; Haglund, A.; and Runeson, B. 2021. Columbia-Suicide Severity Rating Scale Screen Version: initial screening for suicide risk in a psychiatric emergency department. *Psychological medicine*, 1–9.

Gaur, M.; Aribandi, V.; Alambo, A.; Kursuncu, U.; Thirunarayan, K.; Beich, J.; Pathak, J.; and Sheth, A. 2021. Characterization of time-variant and time-invariant assessment of suicidality on Reddit using C-SSRS. *PloS one*, 16(5): e0250448.

Gupta, S.; Agarwal, A.; Gaur, M.; Roy, K.; Narayanan, V.; Kumaraguru, P.; and Sheth, A. 2022. Learning to Automate Follow-up Question Generation using Process Knowledge for Depression Triage on Reddit Posts. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, 137–147.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1): 5485–5551.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Roy, K.; Gaur, M.; Soltani, M.; Rawte, V.; Kalyan, A.; and Sheth, A. 2023. ProKnow: Process knowledge for safety constrained and explainable question generation for mental health diagnostic assistance. *Frontiers in Big Data*, 5.

Sheth, A.; Gaur, M.; Roy, K.; Venkataraman, R.; and Khandelwal, V. 2022. Process Knowledge-Infused AI: Toward User-Level Explainability, Interpretability, and Safety. *IEEE Internet Computing*, 26(5): 76–84.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zhang, Z.; Han, X.; Liu, Z.; Jiang, X.; Sun, M.; and Liu, Q. 2019. ERNIE: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129*.