WISE Causal Models: Wisdom Infused Semantics Enhanced Causal Models - A Study in Suicidality Diagnosis

Anonymous Author(s)
Affiliation
Address
email

Abstract

The COVID-19 Pandemic has highlighted the gap between the number of mental health care seekers and care providers. Netizens have taken to internet-based platforms such as Reddit to express their experiences. Mental illness diagnosis processes have clinically accepted causal interpretations and semantics. Curiously, mental illness diagnosis accuracy is low relative to similar well-studied illnesses. Motivated by this discrepancy, we propose *Wisdom Infused Semantics Enhanced* (WISE) causal models, inspired by the wisdom of the crowd idea that learns from a collective agreement among causal models and their semantics for mental illness diagnoses. We use suicidality diagnosis task descriptions, datasets, and baseline methods to evaluate the effectiveness of WISE causal models. Our experiments show that learning WISE causal models improve performance on these tasks.

1 The Wisdom of the Crowd

The wisdom of the crowd is the collective opinion of a diverse group of individuals as a proxy for expert wisdom [1]. Consequently, if each individual lays out a causal model for the task to solve, one might expect an aggregated vote for actual causation to be of a higher confidence than a single expert causal model. Establishing causality this way is similar to the idea behind the Central Limit Theorem (CLT). The CLT states that the sum of independent random variables (RVs) tends toward a normal distribution [2]. Also, as the number of measurements from the individual RVs reaches large numbers, the effects of noise, measurement artifacts, and confounding variables are "averaged out".

The causal influences approximated as a tree-structure to illustrate context-specificity shown in Figure 1 (a) are expert hypotheses generally accepted among the community [3]. However, suicidality is a peculiar disease in two ways. (1) The accuracy of diagnosis is significantly lower than in similarly well-studied diseases such as diabetes or hypertension. (2) The causal directions are atypical. For example, consider the causal order in the example Post X2 in Figure 1 (c). The post matches concept 1 - Wish to be dead. Thus, it is apparent that the symptom "wish to be dead" caused this person's disease of suicidal ideation. Contrast this with the canonical disease model, where the illness causes symptoms that, in turn, cause the observations that we see and bring to the doctor's attention. To motivate our work, we assume that peculiarity (2) may be causing peculiarity (1). Under our assumption, medical experts may need to postulate other causal models beyond those shown in Figure 1 (a). New causal models arising from such an effort might account for the relatively low suicidality diagnosis accuracies. We propose *Wisdom Infused Semantics Enhanced* (WISE) causal models, inspired by the wisdom of the crowd idea that learns from a collective agreement among causal models and their semantics for suicidality diagnosis. Our experiments show that WISE causal models improves suicdality diagnosis accuracy using baseline implementations on a well-studied suicidality diagnosis task.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

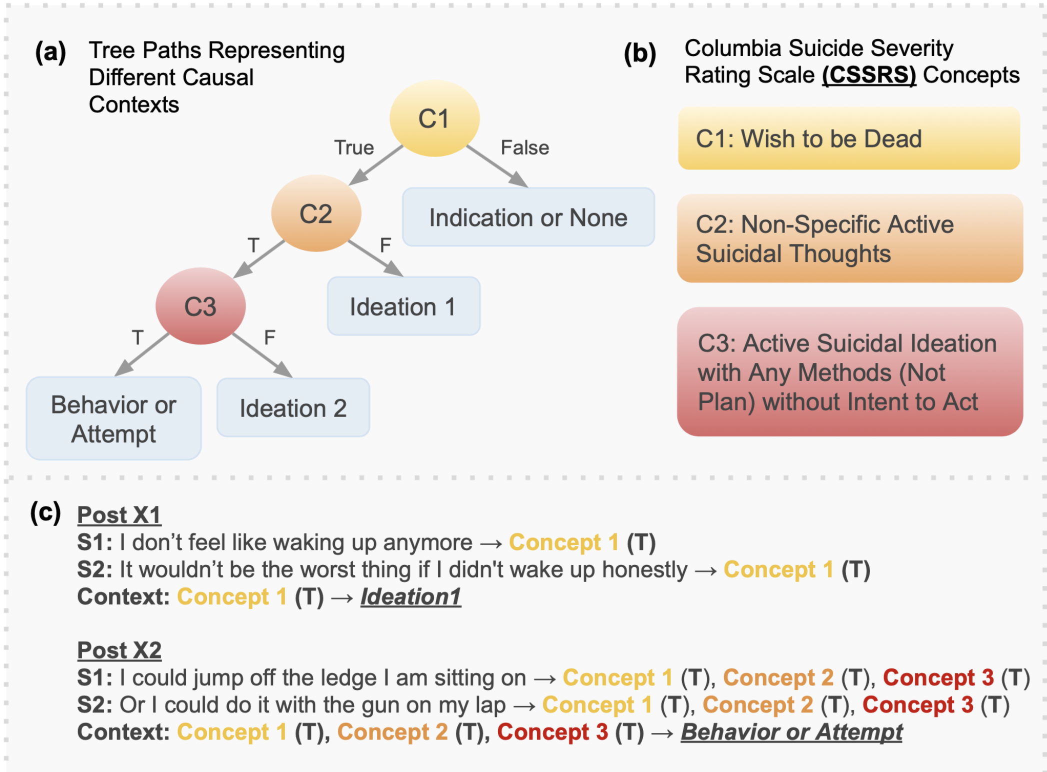

Figure 1: **(a)** Shows the domain expert consensus-based causal model approximated as a tree. The tree paths capture varying contexts that can cause suicidality [3]. **(b)** Shows the descriptions of the suicidality concepts using standard medical vocabulary. **(c)** Shows example Reddit posts annotated with suicidality concepts at the sentence level. A tree path from the root to a leaf determines the suicidality label for that post.

2 Task Description and Data

We choose the suicidality diagnosis task by Gaur et al. to test the performance of WISE causal models[4]. We denote this dataset as MHDA. The data contains high-quality expert annotations on Reddit posts from suicide-related subreddits. The annotation method ensures minimal noise from measurement artifacts and high agreement among the expert annotators. We use the expert consensus causal diagnosis model information in Figure 1 (a) for the causal contexts. Figure 1 (a, b) illustrate the various contexts (each tree path represents a context) under which suicidality diagnosis causal outcomes can arise. The task is to predict the the suicdality diagnosis outcomes, namely - *indication*, *ideation1*, *ideation2*, and *behavior or attempt*. Figure 1 (c) Shows example Reddit posts annotated with suicidality concepts at the sentence level. We denote the augmented dataset as k-MHDA. We defer construction details of k-MHDA from the causal model in Figure 1 (a) and the MHDA data to the appendix Section Constructing k-MHDA as it is not the main focus of the paper[1].

3 WISE Causal Models

We perform causal model searches using randomized subsets of posts sampled from the k-MHDA dataset. We make the following prior assumptions for our baseline implementation. Alterations to our assumptions provide possible directions for future work.

[1]We will release the k-MHDA dataset along with code to construct it.

Assumption 1: The true causal model underlying the k-MHDA dataset patterns is tree structured *Reality*: The true underlying causal model for the k-MHDA dataset might assume any Directed Acyclic Graph (DAG) structure.

Assumption 2: All posts in the k-MHDA dataset are independent and identically distributed *Reality*: The posts in the dataset may be temporally or otherwise correlated.

Assumption 3: The RVs (concept and leaf variables) in Figure 1 (a,b) are the only RVs that determine the causes of suicidality *Reality*: Confounding RVs may influence the causes of suicidality.

3.1 WISE Causal Model Learning

Under Section 3 Assumption 1, we use the Chow-Liu algorithm to learn tree-structured causal models over the randomized post subsets. Under Section 3 Assumption 2, the set of RVs that determine the true underlying causal model are the set of concepts and suicidality outcomes in Figure 1 (a,b). Under Section 3 Assumption 3, we do not learn causal edges between posts in different randomized subsets.

3.1.1 RV Satisfiability

Concept RV satisfiability: For the suicidality concepts in Figure 1 (b), RV satisfiability is determined by Equation 1 that computes the presence or absence of concepts in post sentences using cosine similarity between the texts. The term x_{sub} is a sentence from the input post x (see Figure 1 (c)) and x_{sub}^R is the representation of the sentence. The terms q_i are the concepts in the Figure 1 (b) and q_i^R are their representations. We use the sentence transformer model by Reimers et al. for representation [5]. The notation cos_sim stands for cosine similarity and the sum $\sum_{x_{sub} \in x}(.) \geq 0.5$ is the algebraic form of the $\vee$ operation. We use this term to reflect that we determine the presence of concept q_i in post x, if any of the sentences $x_{sub} \in x$ contain the concept q_i (see Figure 1 (c)).

$$\sum_{x_{sub} \in x} \left(cos\_sim(x_{sub}^R, q_i^R) \geq \theta_i \right) \geq 0.5 \tag{1}$$

Suicidality Outcome Satisfiability: We evaluate a path from the root to a leaf that determines the outcome for suicidality outcome satisfiability. The outcomes are Indication or None, Ideation1, Ideation2, and Behavior or Attempt (see Figure 1 (a,b)). Branching on the concept satisfiability for posts is done using Equation 1.

Algorithm 1 shows the pseudocode for WISE Causal Mode Learning. The wisdom of the crowd inspires the WISE model outcome - The inferred causal outcome is the expected outcome computed by averaging over K causal model outcomes (see Section The Wisdom of the Crowd). Algorithm 2 shows the WISE inference method for an input post x.

Algorithm 1 WISE Causal Model Learning

Initialize the hyperparameters $\{\theta_i\}, K$ ▷ See Section Experiments and Analysis
Initialize Model Scores Placeholder for K Models: $\{\mathcal{M}_k\}$
for $k \leftarrow 1\ to\ K$ **do**
    $\mathbf{P}_{sub} \sim random$(k-MHDA) ▷ randomized post subset from the k-MHDA posts
    $\{RVs\}$ = computed set of values for the RVs ▷ see Section RV Satisfiability
    $\mathcal{M}_k = Chow\_Liu(P_{sub}, \{RVs\})$ ▷ Learn Chow Liu Tree [6]
    for $(x, y) \in \mathbf{P}_{sub}$ **do**
        $P(y \mid x, \mathcal{M}_k, \{\theta_i\}) \sim \mathcal{M}_k(x, \{\theta_i\})$ ▷ Use Exact or Approximate Inference
    Compute Model Score as $\mathcal{S}_{\mathcal{M}_k} = \prod_{(x,y) \in \mathbf{P}_{sub}} P(y \mid x, \mathcal{M}_k, \{\theta_i\})$.
    Add Model Score $\mathcal{S}_{\mathcal{M}_k}$ to the set $\{\mathcal{M}_k\}$
Return $\{\mathcal{M}_k\}$

Algorithm 2 WISE Causal Model Inference

$P_{WISE}(y \mid x, \mathcal{M}_k, \{\theta_i\}) = \mathbb{E}_{\{\mathcal{M}_k\}} P(y \mid x, \mathcal{M}_k, \{\theta_i\}))$ $\triangleright$ Compute Expectation over $\{\mathcal{M}_k\}$

Return $P_{WISE}(y \mid x, \mathcal{M}_k, \{\theta_i\})$

4 Experiments and Analysis

Figure 2 (a) shows the superior quantitative performance of the WISE causal model over the expert causal tree (ECT) and a model called BLM. BLM is the best performing large language model from BERT, T5, and XLNET, fine-tuned for our task. We see that the simple decision tree ECT model achieves a significant jump in accuracy over the best-performing transformer model, which shows the immense value of utilizing publicly available domain expert knowledge for domain-specific tasks. The WISE causal model accuracy shows that expectation based error-correction does lead to performance improvements over the domain-expert hypothesis. It is unclear, however, what the source of such errors might be.

Figure 2 (b) provides color-coded visualization of the WISE causal model inference program outputs. The color codes help visualize the three concepts at the top of the program output snapshot. In this case, the program finds concepts 1 and 3 to be true in the test post and hence infers the outcome *Behavior or Attempt*. Our visualization shows that for this example, the inference agrees with the tree structure in Figure 1 (a), i.e., the path leading to the inference Behavior or Attempt.

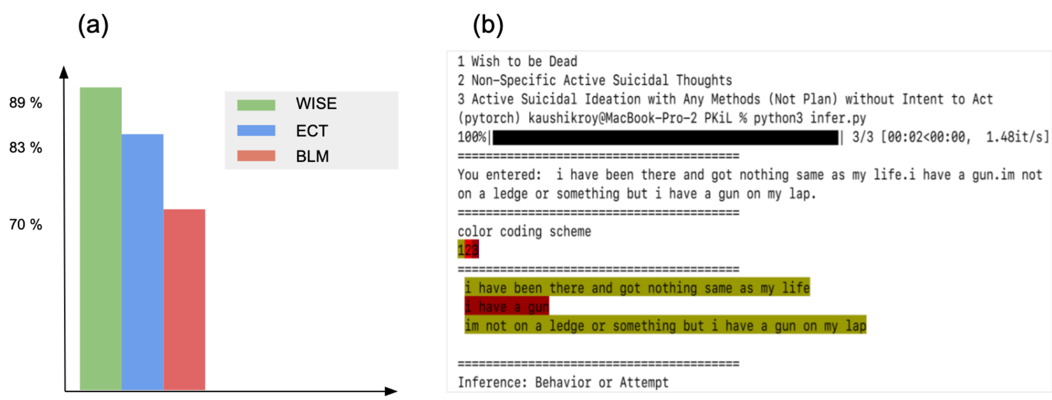

Figure 2: **(a)** Shows how the test-set prediction accuracies of the WISE causal model compared to the expert causal tree (ECT, see Figure 1 (a)). **(b)** Shows how the program provides visualizations to explain the inference using the suicidality concepts for a given test case. The color codings are as per Figure 1 (b).

5 Conclusion and Future Work

We propose WISE causal model learning and inference. Our experiments show that WISE models improve upon data-driven baselines and expert-designed models. Future work on WISE models will involve relaxing the Assumptions in the section WISE Causal Models and measuring the statistical significance of the improvements obtained using WISE models.

References

[1] James Surowiecki. The wisdom of crowds: Why the many are smarter than the few. london. *Abacus*, page 295, 01 2004.

[2] Lucien Le Cam. The central limit theorem around 1935. *Statistical science*, pages 78–91, 1986.

[3] K Posner, D Brent, C Lucas, M Gould, B Stanley, G Brown, P Fisher, J Zelazny, A Burke, MJNY Oquendo, et al. Columbia-suicide severity rating scale (c-ssrs). *New York, NY: Columbia University Medical Center*, 10, 2008.

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

[5] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[6] CKCN Chow and Cong Liu. Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory*, 14(3):462–467, 1968.

A Constructing k-MHDA

There are 500 Reddit posts in the MHDA dataset. Figure 1 (a,b) shows the causal contexts corresponding to suicidality. We can construct a probabilistic decision tree that takes input post x and outputs an outcome y from among the leaves. We can write the tree in algebraic form as shown in **Equation** 2.

$$P(y \mid x, \{\theta_i\}) = \sum_{y \in Outcomes} p_y \prod_{i=1}^{3} \sum_{x_{sub} \in x} \left(cos\_sim\left(x_{sub}^R, q_i^R\right) \geq \theta_i \right) \geq 0.5 \tag{2}$$

p_y is the ground truth probability for each outcome. Index i iterates through the 3 concepts in Figure 1 (b). x_{sub} denotes a sub-fragment of the input post (1 sentence, 2 sentence, etc.). q_i denotes the concept texts from the 3 concepts that i indexes. x_{sub}^R and q_i^R are representations of the post sub-fragment and the concept texts using the sentence-transformer published by Reimers et al. [5].

Equation 3 determines the presence or absence of concept q_i in a post sub-fragment x_{sub}. First, we compute the cosine similarity between their sentence-transformer representations x_{sub}^R and q_i^R. If the resulting value is $\geq \theta_i$, we determine that the concept q_i is present in x_{sub}, else we determine that the concept q_i is absent in x_{sub}.

$\sum_{x_{sub} \in x}(.) \geq 0.5$ in Equation 2 is the algebraic form of the $\vee$ operation as we determine that concept q_i is present in the post x, if any of the post fragments $x_{sub} \in x$ show presence of concept q_i.

$$\left(cos\_sim\left(x_{sub}^R, q_i^R\right) \geq \theta_i \right) \tag{3}$$

We can then evaluate the Bernoulli Loss $\mathcal{L}$ given an input, outcome pair (x, y) and parameters $\{\theta_i\}$ as:

$$\begin{aligned} \mathcal{L}(x, y, \{\theta_i\}) = P(y \mid x, \{\theta_i\}) log(P(y \mid x, \{\theta_i\}))+ \\ (1 - P(y \mid x, \{\theta_i\})) log(1 - P(y \mid x, \{\theta_i\})) \end{aligned} \tag{4}$$

We use grid-search to find a configuration of parameters $\{\theta_i\}$ and post sub-fragment x_{sub} that has the maximum value for $\prod_{(x,y) \in \mathbf{MHDA}} \mathcal{L}(x, y, \{\theta_i\})$. We vary each individual θ_i in the range -1 to 1 (the *range of the cosine function*) and x_{sub} takes values from the set $\{1, 2, 3\}$.

Inference is carried out as it is in a decision tree classifier with the concept presence or absence at each branch, evaluated using **Equation** 3.

Causal Context Annotation with outputs from grid-search: The grid-search yielded outputs $\{\theta_i\} = \{0.3, 0.5, 0.3\}$, and post sub-fragment size $|x_{sub}| = 1$ (one sentence). Therefore the post "I don't feel like waking up and have a gun. Oh well." is annotated with the causal context: **(Concept 1 (T)), Concept 2 (T), Concept 3 (T) =** *Behavior or Attempt*, as evaluation of **Equation** 3 determines absence of **Concept 1**, **Concept 2**, and **Concept 3** in the post sentence"I don't feel like waking up and have a gun". The evaluation uses the grid-search outputs of $\{\theta_i\}$. The second sentence "Oh well' is not necessary to evaluate as we determine a concept's presence or absence in the post if any of the post fragments x_{sub} (one sentence) show the presence of the concept.