Inference Gap in Domain Expertise and Machine Intelligence in Named Entity Recognition: Creation of and Insights from a Substance Use-related Dataset

Sumon Kanti Dey[†,1], Jeanne M. Powell [1], Azra Ismail [1], Jeanmarie Perrone [2], Abeed Sarker [1]
[1] *Department of Biomedical Informatics, Emory University, Atlanta, GA 30322, USA*
[2] *Department of Emergency Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA*
[†] *E-mail: sumon.kanti.dey@emory.edu*

Appendix A. Limitation of the previous study

Despite this importance and potential, relatively few studies have focused on the first-person experiences of individuals who misuse opioids. A notable effort in this area is the work by Ge et al. [1], which attempted to annotate and extract opioid-related impacts from Reddit posts. A notable effort in this area is the work by Ge et al. [1], which attempted to annotate and extract opioid-related impacts from Reddit posts. Upon closer examination, we identified multiple sources of annotation noise and inconsistency that limit the dataset's utility for fine-grained, first-person impact detection. One major issue is the inclusion of entity spans from second-person and third-person perspectives. For example, the sentence "You can taper as slow as possible and drop in the tiniest increments and still feel withdrawal because the alkaloid concentration varies so much ..." was annotated as a clinical impact, despite being framed in the second person with no clear reference to the narrator's own experience. Similarly, the sentence "My mother struggled with mobility" was labeled as a clinical impact, although it describes a third-person experience. While such narratives may carry relevance in broader discourse analysis, our work is specifically concerned with identifying self-reported social and clinical impacts—where the speaker directly references their own condition or circumstance. This distinction is particularly important for applications such as self-reported symptom tracking and personalized care modeling.

We also found several cases of overgeneralized annotations where entire sentences were marked as impact entities, even though only a subset of words represented the actual impact. For instance, in the sentence "So I did 45 days in a rehab here in Michigan," the word "rehab" represents the clinical impact, yet the full sentence was annotated as such. Similarly, "I was homeless because of drugs" was labeled entirely as a social impact, although only the word "homeless" denotes the impact. Another example includes the sentence "I go to Alcoholics Anonymous, but I need more help/support," where only "Alcoholics Anonymous" should be identified as a social impact, but the whole sentence was tagged. In some instances, annotations were simply incorrect; for example, the sentence "I had to wait like 38 hours, and it finally worked" was labeled as a social impact, even though it does not convey any social or clinical consequence. Furthermore, neutral mentions of substances such as "methadone," "kratom", "Heroin", and "drugs" were often annotated as clinical impacts, despite not indicating any actual impact or personal consequence in the context of those sentences. Annotation

inconsistencies have been in Table A1.

These issues introduce noise and reduce the effectiveness of machine learning models trained on the dataset, particularly transformer-based models, which have shown poor performance——reportedly achieving zero F1 scores——under these conditions.

Table A1: Comparison of Original and Refined Annotations for Clinical and Social Impact Detection

Tokens	I	was	a	homeless	junkie	on	the	streets	of	Florida	.
Original Annotation	SocialImpacts								_	_	_
Refined Annotation	_	_	_	SocialImpacts		_	_	_	_	_	_

Tokens	So	I	did	45	days	in	a	rehab	here	in	michigan .
Original Annotation	ClinicalImpacts								_	_	_
Refined Annotation	_	_	_	_	_	_	_	ClinicalImpacts	_	_	_

Appendix B. Conditional Random Fields (CRF)

Conditional Random Fields (CRFs) [2] are probabilistic models widely used for sequence labeling tasks such as Named Entity Recognition (NER). CRFs capture the dependencies between adjacent labels in a sequence, making them particularly effective in domains where contextual consistency is critical, such as social and clinical impact detection. Unlike models that assume independent label predictions, CRFs consider both the current token and its surrounding label structure, reducing inconsistent or invalid tagging patterns.

Let $x = \langle x_1, x_2, \ldots, x_n \rangle$ be an input sequence of tokens and $y = \langle y_1, y_2, \ldots, y_n \rangle$ be the corresponding sequence of labels. A linear-chain CRF defines the conditional probability of a label sequence given the input as:

$$P(y \mid x) = \frac{1}{Z_x} \exp\left(\sum_{t=1}^{n} \sum_{k=1}^{K} \lambda_k f_k(y_{t-1}, y_t, x, t)\right)$$

where Z_x is a normalization factor that sums over all possible label sequences, ensuring a valid probability distribution. Each $f_k(y_{t-1}, y_t, x, t)$ is a feature function that encodes a specific combination of the previous label y_{t-1}, the current label y_t, the observation sequence x, and the current position t. These functions may represent transition features (e.g., from one label to another) or emission features (e.g., current word shape or embedding values). The parameters λ_k are learned weights that indicate the importance or reliability of each feature function during training. The label y_t corresponds to the class assigned to the word x_t at time step t, while y_{t-1} represents the label of the preceding token.

By learning both emission and transition dynamics, CRFs reduce label ambiguity and enforce coherent tag sequences across the input. In this study, the CRF layer is applied on

top of contextualized token embeddings derived from a transformer encoder, enabling robust recognition of entities related to social and clinical impacts.

Appendix C. Annotation Guidelines

We annotated entities belonging to two categories–*SocialImpacts* and *ClinicalImpacts*– in first-person narratives of opioid misuse. The following guidelines were applied to ensure consistency, reliability, and alignment with the intended task objectives.

General Instructions

- Annotations capture only meaningful, self-reported consequences of opioid misuse expressed by the individual.
- Both *SocialImpacts* (e.g., job loss, family disruption) and *ClinicalImpacts* (e.g., withdrawal, depression, hospitalization) are included.
- All annotations must reflect the individual's own lived experience.

Inclusion Criteria

Entities were annotated when they met all of the following criteria:

- **First-person account:** A social or clinical impact was annotated only if it was described in a first-person account directly related to the poster.
  Example: "I lost my job" → annotated; "My brother lost his job" → not annotated.
- **Ambiguous context (assumed impact):** When opioid involvement could not be ruled out, the impact was assumed to be related.
  Example: "It caused me to fight with my family", where "fight with my family" annotated as *SociallImpacts*.
- **Polysubstance mention:** If a post mentioned multiple substances, annotate assuming opioid misuse contributed.
  Example: "I abuse alcohol and heroin, which has affected my health" → annotated as *ClinicalImpacts*.
- **Mental health symptoms:** Mental health issues were annotated as *ClinicalImpacts* unless explicitly attributed to another cause.
  Included: "I feel depressed all the time."
  Excluded: "We broke up, so I am sad.", where sad is clearly linked to the breakup and not opioid use
- **Care-seeking behavior:** Mentions of rehab, counseling, or treatment were annotated as *ClinicalImpacts*.
  Example: "I went to rehab last month" → "went to rehab" is annotated.

Exclusion Criteria

The following were explicitly excluded:

- **Third-person accounts:** Impacts involving friends, family, or others.
  Example: "My brother lost his job" → not annotated.

- **Drug names:** Mentions of specific drugs were not annotated as impacts (captured separately).
- **Personal pronouns in spans:** Personal pronouns (e.g., "I", "he", "she", "they", "my", "our") were excluded from the annotated span if they did not contribute directly to the social or clinical impact.
  Example: "I lost my job" → span: *lost my job*.
- **Modifiers in spans:** Temporal references, adjectives, adverbs, and similar modifiers were excluded unless integral to meaning.
  Example: "I am feeling really tired and crummy" → span: *tired and crummy*.

Appendix D. Fine-tuning and LLM Prompting Details

Appendix D.1. *Language Model Fine-tuning*

We fine-tuned several encoder-based Pretrained Language Models (PLMs) on our annotated dataset, experimenting with various hyperparameters to identify optimal configurations for the Named Entity Recognition (NER) task. We explored multiple combinations of learning rates, batch sizes, and other critical hyperparameters, as well as configurations both with and without a Conditional Random Field (CRF) layer. Table D1 summarizes the ranges of key hyperparameters examined during model training.

Table D1: Fine-tuning hyperparameter ranges explored across multiple encoder-based pre-trained language models (PLMs), with and without a Conditional Random Field (CRF) layer.

Hyperparameter	**Values Explored**
Base Model	Multiple PLMs (encoder-based)
Batch Size	{8, 16, 32}
Learning Rate	{1e-5, 2e-5, 3e-5, 5e-5}
Dropout	{0.2, 0.3}
Number of Epochs	{7,10}
Weight Decay	0.01
Learning Rate Scheduler	Linear
Gradient Clipping (max grad norm)	1.0
Early Stopping Patience	3
Conditional Random Field (CRF)	{True, False}

Appendix D.2. *LLM Prompting Details*

For evaluating LLMs such as GPT-4o, LLaMA3-70B, and Gemini, we utilized the structured prompt template described in Table D2. To ensure consistency and reproducibility across all LLM-based experiments, we set the temperature to 0.2, emphasizing deterministic and reliable outputs.

Table D2: Prompt used for in-context learning. Few-shot examples ($n = 3$ or $n = 5$) are dynamically selected based on semantic similarity to the input.

You are a medical AI assistant that classifies tokens in a Reddit post into the following categories:

ClinicalImpacts: Health and well-being effects.
SocialImpacts: Societal or community-level effects.
O: Tokens outside these categories.

Strict Annotation Rules
1. Annotate **ONLY** first-person experiences. Ignore third-party reports.
2. Label all drug names (e.g., `heroin`, `fentanyl`) as `O`.
3. Label personal pronouns (e.g., `I`, `my`) as `O` – they are not part of entity spans.
4. ASSUME opioid involvement unless a non-opioid cause is clearly stated.
5. If multiple substances are mentioned, default to opioid-related impact when unsure.
6. Label mental health terms (e.g., `depression`) as `ClinicalImpacts` unless context clearly shows a non-opioid cause.
7. Label non-integral words (e.g., adjectives, adverbs, or temporal words like `very`, `suddenly`) as `O` if they are not essential to the entity span.
8. Corrupted or unreadable tokens (e.g., `Ìm`, `?`, `##`) must be labeled as `O`.
9. Maintain the exact token order and label all tokens.
10. If unsure about a token, label it as `O`.

Output Format
Return token-label pairs in the following format:
`token-Label token-Label token-Label ...`

Few-shot Examples (Top-N, dynamically retrieved):
Example 1: {token-Label token-Label token-Label ...}
...
Example 3: {token-Label token-Label token-Label ...}
...

New Input:
`Tokens: [token_1, token_2, ..., token_n]`
Output:

References

[1] Y. Ge, S. Das, K. O'Connor, M. A. Al-Garadi, G. Gonzalez-Hernandez and A. Sarker, Reddit-impacts: A named entity recognition dataset for analyzing clinical and social effects of substance use derived from social media, *arXiv preprint arXiv:2405.06145* (2024).

[2] J. Lafferty, A. McCallum, F. Pereira *et al.*, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in *Icml*, (2)2001.