RACAI at #SMM4H-HeaRD: Named Entity Recognition for Detecting the Impacts of Drug Abuse in Social Media Posts: Zero-Shot and Fine-Tuning Approaches

Tiberiu Boros and Radu Chivereanu
Research Institute for Artificial Intelligence "Mihai Draganescu",
Romanian Academy
Calea 13 Septembrie, Bucharest, Romania
tibi@racai.ro, radu.chivereanu@racai.ro

Abstract

In this work, we address the detection of drug abuse repercussions in Reddit posts, as part of SMM4H-HeaRD Task 7: Extraction of Social and Clinical Impacts of Substance Use from Social Media Posts. We evaluate multiple approaches, including fine-tuning and zero-shot inference, across several deep learning architectures. Our best result is obtained using an adapter-based fine-tuning approach on the DeBERTaV3 model. In addition, we explore text-based evolutionary optimization for Gemma 4 workflows and show that, on this task, they achieve competitive performance with the supervised DeBERTaV3 setup.

1 Introduction

The non-clinical use of controlled substances, particularly opioids, remains a public health crisis. Clinical records provide good insights, but social media posts provide a unique view into people's lives, revealing what the written record fails to capture: **the full spectrum of consequences experienced by individuals**.

This paper describes our approach to the SMM4H-HeaRD 2026 Shared Task (Lopez-Garcia et al., 2026). The task addresses the extraction of social and clinical impacts of substance abuse from social media narratives, more precisely opioid misuse. Particularly challenging are the informal nature of social media posts, the use of **slang** to mask explicit substance usage, and the limited amount of available data. Our contributions are two-fold:

(i) First, we provide a comparison of both supervised and zero-shot approaches for the task, evaluating architectures such as DeBERTaV3 (He et al., 2023), Gemma 4 (Gemma Team, Google DeepMind, 2026), and GliNER (Zaratiana et al., 2024).

(ii) Second, we conduct a data and error analysis, highlighting the main challenges associated with identifying nuanced social and clinical impact mentions in noisy user-generated content.

Our submission for the evaluation phase, an adapter-based fine-tuned DeBERTaV3 model, achieved the best Relaxed F1 score in the competition, reaching 0.61. **Code is available here**[1].

2 Related Work

Named entity recognition (NER) is one of the well-established tasks of natural language processing (NLP). The literature is vast for this domain, starting from linear classifiers (Takeuchi and Collier, 2002; Makino et al., 2002) with feature engineering and working toward deep learning approaches (Hakala and Pyysalo, 2019; Zeng et al., 2017). Low-resource NER is a niche covered by several works (Yohannes and Amagasa, 2022; Zaratiana et al., 2024), usually through unsupervised or semi-supervised approaches or through downstream adaptation of pretrained language models.

The specific challenge of extracting clinical and social impacts from opioid-related social media posts was studied by Dey et al. (2025), who found that fine-tuned encoder models consistently outperform prompted Large Language Models (LLMs) on the task. This *inference gap* is further documented in broader clinical NER literature: prompted LLMs exhibit larger distributional discrepancies from gold annotations than fine-tuned models, including in predicted span length (Zhao and Goto, 2025), and it can be partially reduced by spending more effort in prompt engineering (Hu et al., 2024).

Interesting to the present work is the use of prompt-tuning techniques for zero-shot or few-shot adaptation of LLMs to downstream tasks (Tong et al., 2025). We experiment with a modified version of the text-based evolutionary optimization method—Genetic Pareto (GEPA) (Agrawal et al., 2025)—to automatically develop a complete LLM-

[1] https://github.com/racai-ai/hner

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 121–126

based NER workflow. Our results indicate that this can narrow the inference gap considerably on the impact extraction task, being competitive with the fine-tuned encoder performance on the Relaxed F1 metric.

3 Data Analysis

The problem is framed as a named entity recognition (NER) task, where participants have to determine the boundaries of the two entity types (social and clinical impacts). The provided dataset is composed of sentences tokenized at the word level and associated entity spans, labeled using the Beginning-Inside-Outside (BIO) standard. Data is partitioned into a training set, a development set, and a test set, the latter being hidden during the development phase of the Shared Task.

A particular feature of this dataset is that only "**first-person experiences**" are taken into consideration. Indirectly, this means the detection of clinical and social impacts using local-only or limited window classification is inefficient, since the classifier requires a global state indicating whether the narration is from a first-person perspective. Additionally, posts in the training and development set are split into multiple examples.

From a quantitative perspective, the training set contains a total of 842 examples, with 173 unique clinical-impact entities (256 total occurrences) and 78 unique social-impact entities (87 total occurrences). The development set is composed of 258 samples, with 75 unique clinical-impact entities (92 total occurrences) and 27 unique social-impact entities (27 total occurrences). Also, a large number of sentences contain no labels.

3.1 Training and Development Set Analysis

While formulated as a NER task, there are several challenges, aside from data availability, that complicate this task:

(i) **Language flexibility and slang** with which substance abuse, clinical and social impacts can be expressed. In contrast, standard NER involving persons, locations, organizations, etc. has strong textual and lexicographic clues (e.g. capitalized letters - "New York", standard wording - "I live in Lisbon" etc.)

(ii) **Causality**: There are two layers of causality involved: (a) the "first person" condition and (b) the requirement to label social and clinical impacts only if they are an implication of substance abuse (e.g. losing one's job due to budget cuts does not qualify as a social impact, because it is not a consequence of substance abuse). According to guidelines, in ambiguous contexts, if opioid involvement cannot be ruled out, it is considered **related**.

Description	Strict	Fuzzy
Clinical impacts		
Overlapping entities	19	33
Only in trainset	154	140
Only in devset	56	9
Social impacts		
Overlapping entities	2	5
Only in trainset	76	73
Only in devset	25	24

Table 1: Data overlap between train and dev sets based on identical and fuzzy-matched entities.

The overlap between the training and development set entities is small, as outlined in Table 1. Table 2 contains examples.

4 System Architecture and Methods

We evaluate four approaches:

(i) **BiLSTM baseline** using word-level and morphological features (Section 4.1);

(ii) **DeBERTaV3 models** via full fine-tuning or adapter-based fine-tuning (Section 4.2.1);

(iii) **GLiNER for zero-shot NER** using a tiered tagging strategy (Section 4.2.2);

(iv) **LLMs** with automatic prompt tuning (Section 4.3).

For all weight updates, we use a weighted sum of cross-entropy ($\mathcal{L}_{CE}$) and Dice loss ($\mathcal{L}_{D}$) (Li et al., 2020) to **mitigate class imbalance**. Each section includes a discussion of approach-specific findings.

4.1 Baseline model

Our baseline is a token-level **BiLSTM** using a tripartite input representation $I_w = E_w + H_w + M_w$ (Dozat et al., 2017), summing: (1) precomputed word embeddings (Mikolov et al., 2013), (2) trainable word embeddings (freq $\geq$ 2), and (3) morphological embeddings (Bojanowski et al., 2017). The latter (M_w) averages trainable vectors of word-internal n-grams ($n \in \{3, 4, 5\}$) and boundary markers.

To reduce overfitting, we apply a dropout strategy (Dozat et al., 2017) that masks full embedding vectors and rescales the sum proportionally (e.g., $\times 2$ or $\times 3$). This encourages a unified embedding space and forces the model to leverage context during training.

Training set	Development set	#t	#d	Training set	Development set	#t	#d
"drug-induced psychotic break"	"psychotic break"	1	1	"sleep issues"	"issues"	1	1
"drug induced psychosis"	"meth induced psychosis"	2	1	"therapy"	"therapy part"	2	1
"feeling worse"	"worse"	1	1	"pitted acne"	"acne"	1	1
"Outpatient"	"outpatient clinic"	1	1	"lost my truck"	"lost my job"	1	1
"precipitated withdrawal"	"withdrawal"	1	7	"got addicted"	"addicted"	1	3
"physically addicted"	"addicted"	1	3	"clinic"	"outpatient clinic"	2	1
"rehab systems"	"rehab"	1	2	"addiction counseling"	"addiction"	1	1
"lost my family"	"lost my job"	1	1	"I seriously lost everything that I cared about other than my family"	"lost everything I loved in life other than my family"	1	1

Table 2: Aligned entities using fuzzy matching and their frequencies in both datasets. #t represents the frequency in the training set and #d in the development set.

```
[ENT] SocialImpacts [ENT] ClinicalImpacts [SEP]
                    +
          This is just an example .
```

Figure 1: Example input for the GLiNER approach.

Empirically, this was the only configuration where a weighted sum of $\mathcal{L}_D$ and $\mathcal{L}_{CE}$ outperformed $\mathcal{L}_{CE}$ alone; without the tripartite dropout, the model overfit the training data significantly.

4.2 Encoder-Based Models

4.2.1 Fine-tuning and Adapters

We utilize DeBERTaV3-Large as our backbone, aligning the tokenized sentences with the model's native tokenizer. Following empirical validation, we train the model to predict labels only for the "head" (first sub-token) of each word, as sub-word labeling proved suboptimal.

We compare full fine-tuning against Parameter-Efficient Fine-Tuning (PEFT) using Adapters (Pfeiffer et al., 2020).

Adapters insert bottleneck layers between transformer blocks, enabling parameter-efficient adaptation while maintaining—and in some cases exceeding (Pfeiffer et al., 2021)—full fine-tuning performance.

After performing a grid search for the weights of the $\mathcal{L}_{CE}$ and $\mathcal{L}_D$ combination, $\mathcal{L}_{CE}$ alone produced the best results for this architecture.

4.2.2 Zero-shot encoder architectures: GLiNER

GLiNER enables zero-shot NER by scoring the similarity between token-level embeddings and label embeddings. This approach relies on BERT-like models and uses the prompt prefixing strategy. More precisely, the target token classes are prefixed to the input sentence using special tokens (Figure 1). A neural network built on top of the encoder output is trained to align the embeddings of the target classes with those of the tokens in the input segment that belong to those classes. This works for fine-grained classes, but coarse classification (e.g. Clinical or Social impacts) underperformed. To improve accuracy, we decomposed the meta-classes into granular sub-classes:

ClinicalImpacts: symptom, withdrawal, addiction, substance, drug use, abuse, overdose, treatment, rehabilitation, mental health.

SocialImpacts: homelessness, legal issues, arrest, imprisonment, relationship breakdown, family estrangement, social isolation, job loss, financial hardship.

While GLiNER supports fine-tuning by updating the underlying model, we focused exclusively on zero-shot performance.

4.3 Large Language Models (LLMs)

Based on performance and open-source accessibility, we evaluate Gemma 4 31B-it. To circumvent the bias and effort of manual prompt engineering, we employ two automatic optimization techniques to evolve an end-to-end NER workflow in Python: (i) a custom genetic iterative approach and (ii) **Genetic-Pareto (GEPA)** (Agrawal et al., 2025).

Optimization Process: Both approaches evolve candidates through iterative evaluation and LLM reflection. A candidate represents a script that, given text, extracts the required entities using one or multiple LLM calls. Candidates are tested in 8-sample batches; for the custom method, we select candidates that improve either Relaxed or Strict

Method	C-Impacts		S-Impacts		Dev		Test	
	P	R	P	R	S-F1	R-F1	S-F1	R-F1
Baseline BiLSTM	0.59	0.28	0.08	0.07	0.25	0.30	*	*
GliNER	0.25	0.20	0.64	0.08	0.17	0.19	*	*
DeBERTaV3 fine-tuned	0.73	**0.54**	0.60	0.38	0.44	0.56	*	*
DeBERTaV3 adapters	**0.81**	0.45	0.63	0.44	0.44	0.55	**0.45**	**0.61**
Gemma 4 31B - Manual Prompting	0.66	0.53	0.69	**0.46**	0.35	**0.57**	*	*
Gemma 4 31B - Iterative	0.72	0.53	**0.73**	0.39	**0.45**	**0.57**	*	*

Table 3: Model performance displaying relaxed precision and recall per class, alongside global F1-scores (R-F1: Relaxed F1, S-F1: Strict F1) on dev and test sets.

F1, against the full dev set; for the GEPA variant, we optimize Precision and Recall since, **at equilibrium, they approximate a Pareto frontier**, by marginally improving F1. We executed the pipeline in two phases:

Seedless Run: Candidates are proposed solely from evaluation feedback and a minimal task description.

Seeded Run: We restart from the best candidate, manually adding the "first-person experience" rule, which the model failed to infer from the data.

The optimal candidate for the custom method achieved a 0.45 strict F1 score and a 0.57 Relaxed F1 on the dev set, using a two-stage architecture: (1) **Span Extraction**, to identify entity text, and (2) **Token Alignment**, to map those spans back to the original indices. The GEPA approach yielded similar results. There were no statistically relevant differences between the two approaches.

4.4 Data augmentation experiments

We explored four synthetic augmentation strategies to address the low-resource nature of the dataset:

(i) Negation/Perspective Shift: Altering the narrative person (e.g., first to third person) and removing labels to reinforce the "first-person" encounter rule.

(ii) Synonym Replacement: Substituting tokens with semantically similar terms.

(iii) Expression Mutation: Transforming simple entity mentions into more complex syntactic structures.

(iv) Synthetic Generation: Using LLMs to create entirely new Reddit-style posts. To increase variability, we incorporated excerpts from real social media posts into the prompt.

Despite the generated data appearing qualitatively sound, none of these techniques improved performance; rather, they degraded it. While a more extensive error analysis is required, we hypothesize that synthetic data introduced noise that shifted the distribution of the data.

5 Experimental validation and discussion

The task proposes the relaxed token-level F1-score as a metric to award partial credit for overlapping boundary predictions (Dey et al., 2025).

Table 3 summarizes model performance across the development set. Our best-performing configuration (DeBERTaV3-Large with adapters) achieved **a 0.61 Relaxed F1 score on the test set**.

5.1 Error analysis on development set

Observation 1: Among the 76 examples misclassified by DeBERTaV3 and the 68 misclassified by our optimized LLM script, 60 were common.
Observation 2: In the 60 common examples, DeBERTaV3 mislabeled 289 tokens and the script 283. 225 tokens were assigned the same incorrect class by both approaches. The overlap may stem from annotation noise and limited context.
Observation 3: The average entity span length is 4.1 tokens in the ground truth, compared to 2.32 tokens for DeBERTaV3 predictions and 1.96 for the LLM-based script, correlating with the difference in recall.

6 Conclusions

We described our participation in the SMM4H-HeaRD 2026 Shared Task 7. In our experiments, pretrained models based on the transformer architecture were the most effective. The adapter-based DeBERTaV3-Large model achieved the highest Relaxed F1 score on the official test set. We also explored zero-shot and prompt-optimized LLM approaches, where iteratively refined and self-reflective LLM pipelines achieved competitive results on the development set.

Future work will focus on understanding why the explored augmentation strategies degraded performance and how they can be improved.

References

Lakshya A. Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J. Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alexandros G. Dimakis, Ion Stoica, Daniel Klein, Matei Zaharia, and Omar Khattab. 2025. GEPA: reflective prompt evolution can outperform reinforcement learning. *CoRR*, abs/2507.19457.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Sumon Kanti Dey, Jeanne M Powell, Azra Ismail, Jeanmarie Perrone, and Abeed Sarker. 2025. Inference gap in domain expertise and machine intelligence in named entity recognition: Creation of and insights from a substance use-related dataset. In *Biocomputing 2026: Proceedings of the Pacific Symposium*, pages 12–26. World Scientific.

Timothy Dozat, Peng Qi, and Christopher D Manning. 2017. Stanford's graph-based neural dependency parser at the conll 2017 shared task. In *Proceedings of the CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies*, pages 20–30.

Gemma Team, Google DeepMind. 2026. Gemma 4: Our most capable open models to date. Google DeepMind Blog. Accessed: 2026-04-24.

Kai Hakala and Sampo Pyysalo. 2019. Biomedical named entity recognition with multilingual bert. In *Proceedings of the 5th workshop on BioNLP open shared tasks*, pages 56–61.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, Kirk Roberts, and Hua Xu. 2024. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, 31(9):1812–1820.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2020. Dice loss for data-imbalanced nlp tasks. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 465–476.

Guillermo Lopez-Garcia, Jose Miguel Acitores Cortina, Jacob Berkowitz, Joey Chan, Ganesh Chandrasekar, Sumon Kanti Dey, Ivan Flores Amaro, Fernando Gallego, Lauren Gryboski, Ari Z Klein, Martin Krallinger, Salvador Lima-López, Tomohiro Nishiyama, Lisa Raithel, Ahmad Rezaie Mianroodi, Roland Roller, Judith Rosell, Frank Rudzicz, Abeed Sarker, and 8 others. 2026. Overview of the 11th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Shared Tasks at ACL 2026. In *Proceedings of the 11th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Workshop and Shared Tasks*. Association for Computational Linguistics.

Takaki Makino, Yoshihiro Ohta, Jun'ichi Tsujii, and 1 others. 2002. Tuning support vector machines for biomedical named entity recognition. In *Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain*, pages 1–8.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. Adapterfusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 487–503. Association for Computational Linguistics.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. Adapterhub: A framework for adapting transformers. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 46–54.

Koichi Takeuchi and Nigel Collier. 2002. Use of support vector machines in extended named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Zeliang Tong, Zhuojun Ding, and Wei Wei. 2025. Evoprompt: Evolving prompts for enhanced zero-shot named entity recognition with large language models. In *Proceedings of the 31st international conference on computational linguistics*, pages 5136–5153.

Hailemariam Mehari Yohannes and Toshiyuki Amagasa. 2022. Named-entity recognition for a low-resource language using pre-trained language model. In *Proceedings of the 37th ACM/SIGAPP symposium on applied computing*, pages 837–844.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. Gliner: Generalist model for named entity recognition using bidirectional transformer. 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 5364–5376.

Donghuo Zeng, Chengjie Sun, Lei Lin, and Bingquan Liu. 2017. Lstm-crf for drug-named entity recognition. *Entropy*, 19(6):283.

Yichong Zhao and Susumu Goto. 2025. Can frontier llms replace annotators in biomedical text mining? analyzing challenges and exploring solutions. *Preprint*, arXiv:2503.03261.