Reddit-Impacts: A Named Entity Recognition Dataset for Analyzing Clinical and Social Effects of Substance Use Derived from Social Media

Yao Ge[1], Sudeshna Das[1], Karen O'Connor[2], Mohammed Ali Al-Garadi[3], Graciela Gonzalez-Hernandez[4], and Abeed Sarker[1,5,*]

[1]Department of Biomedical Informatics, School of Medicine, Emory University, Atlanta, GA
[2]DBEI, The Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA
[3]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN
[4]Department of Computational Biomedicine, Cedars-Sinai Medical Center, West Hollywood, CA
[5]Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA

Abstract

Substance use disorders (SUDs) are a growing concern globally, necessitating enhanced understanding of the problem and its trends through data-driven research. Social media are unique and important sources of information about SUDs, particularly since the data in such sources are often generated by people with lived experiences. In this paper, we introduce Reddit-Impacts, a challenging Named Entity Recognition (NER) dataset curated from subreddits dedicated to discussions on prescription and illicit opioids, as well as medications for opioid use disorder. The dataset specifically concentrates on the lesser-studied, yet critically important, aspects of substance use—its clinical and social impacts. We collected data from chosen subreddits using the publicly available Application Programming Interface for Reddit. We manually annotated text spans representing clinical and social impacts reported by people who also reported personal nonmedical use of substances including but not limited to opioids, stimulants and benzodiazepines. Our objective is to create a resource that can enable the development of systems that can automatically detect clinical and social impacts of substance use from text-based social media data. The successful development of such systems may enable us to better understand how nonmedical use of substances affects individual health and societal dynamics, aiding the development of effective public health strategies.

In Reddit-Impacts dataset, we have a total of 1,380 posts, among which 23% contain words or phrases annotated as clinical or social impacts. Specifically, 246 posts include entities annotated as having clinical impacts, and 72 posts are related to social impacts.

In addition to creating the annotated data set, we applied several machine learning models to establish baseline performances. Specifically, we experimented with transformer models like BERT, and RoBERTa, one few-shot learning model DANN by leveraging the full training dataset, and GPT-3.5 by using one-shot learning, for automatic NER of clinical and social impacts. The dataset has been made available through the 2024 SMM4H shared tasks.

1 Introduction

Substance use disorders represent a critical challenge in public health, with both clinical and social consequences impacting individuals and communities worldwide [1, 2]. The pervasive nature of substance use, encompassing both prescription and illicit drugs, necessitates a deeper understanding of

its impacts to inform more effective interventions and preventative measures [3, 4]. This study introduces the REDDIT-IMPACTS dataset, a unique corpus derived from Reddit, a platform known for its rich, anonymized discussions among diverse groups, including individuals who use drugs [5–7]. The dataset includes posts from 14 opioid-related subreddits, capturing a broad spectrum of experiences and discussions related to substance use.

Our research specifically focuses on the clinical and social impacts of nonmedical substance use. These impacts are critical yet under-represented in the available data, making them an ideal focus for applying few-shot learning techniques to improve named entity recognition (NER) tasks. The clinical impacts encompass the direct effects on an individual's health, while social impacts involve the broader consequences on relationships, communities, and societal structures. While there is an abundance of such information on Reddit, they are embedded in vast volumes of other unrelated information, making it extremely challenging to detect them automatically with high accuracy from naturally distributed data.

This paper details the creation of the REDDIT-IMPACTS dataset, describes our annotation process, provides data statistics, and discusses the application of supervised learning algorithms aimed at enhancing the detection and classification of clinical and social impacts. We employed two models to benchmark the performance of our few-shot learning approach: one utilizing the full training dataset and another leveraging one-shot capabilities of large language models (LLMs) like Generative Pre-trained Transformer (GPT [8, 9]). These models provide baseline performance metrics that can be used for comparative evaluation of future systems customized for this complex NER task. In addition, this study explores innovative methodologies in applying few-shot learning to improve data annotation efficiency and model performance in detecting nuanced entity types like clinical and social impacts.

2 Methods

This study was considered to be exempt (category 4; publicly available data) by the Institutional Review Board of Emory University. The overall study can be divided into 4 steps: (1) data collection, (2) manual annotation, (3) creation of the REDDIT-IMPACTS dataset and (4) NER.

2.1 Data collection

Reddit is popular in the broader community of people who use drugs as it offers anonymity, and Reddit has seen rapid growth in its user base over the last several years. Reddit communities have also been found to serve as a means of social support for people who use drugs. We chose Reddit over other social networks or web-based forums such as Twitter, Bluelight, and Discord for several reasons. While all these sources contain information about substance use, the substance use community of Reddit is much larger and has been extensively used in peer-reviewed research related to substance use and emerging substance use trends. Additionally, Reddit threads are also heavily moderated, and posts must follow community-specific rules. Consequently, while these rules restrict some types of information from being posted, they also ensure that the data are reflective of the topical areas and the volume of spam, posts from bots, or irrelevant content is thereby lower. The existence of standard application programming interfaces (APIs) also makes data collection from Reddit relatively straightforward.

To identify potential Redditors (Reddit subscribers) who self-report opioid usage on Reddit, we identified 14 opioid-related subreddits spanning discussions on prescription and illicit opioids, and collected all retrievable posts using the Python-Reddit API Wrapper for Reddit (PRAW).[1]

The choice of these subreddits was based on their topical relevance and high levels of community discussion and engagement. Collection of data from these subreddits was not keyword-based. Instead, the API allowed the retrieval of all publicly posted threads and the associated comments. After retrieving all available posts of the 47,327 Redditors who had posted on the selected sub-reddits, we selected a random sample of these Redditors (N=13,812) and collected each of their past public posts across all subreddits (i.e., their longitudinal timelines), between November 2006 (corresponding to the earliest post available) and March 2019 (corresponding to the last date of data collection).

[1] https://praw.readthedocs.io/en/latest/

2.2 Annotation

From the 13,812 public timelines we collected, we randomly selected 40 Redditors' timelines (i.e., all their posts in different subreddits) for manual review and annotation. This process finally yielded 26,126 posts for annotation. The annotation process was iterative and involved several steps. The posts were manually analysed to develop the annotation guidelines, and then preliminary rounds of annotation were performed. We then discussed the disagreements, and updated the annotation guidelines for further clarity, and the final annotation was performed on a total of 91,601 sentences (2,500,489 tokens).

Due to the complexity of the annotation task, involving many entity types, and large numbers of posts that contained no entities at all, rather than annotating separately and then computing inter-annotator agreement, the data was first annotated by the lead annotator (KO) based on annotation guidelines and reviewed by two members of the study team. Following the annotation of all posts by two subscribers, the annotations were reviewed by the full team, disagreements were resolved via discussion and the annotation guideline was updated. Subsequent annotations were carried out in the same manner, adhering to the annotation guideline. All disagreements were resolved via discussion.

Based on the annotation guidelines we annotated lexical expressions in posts into 30 entity types that are independent of each other. Among them, 10 entity types belong to the basic personal information category, such as Age, Gender, Marital status, Location, Income, etc. 20 entity types related to medication information, such as Medicine intake, Illegal drug use, NMPDU, Method of intake, etc. Figure 1 shows all 30 entity types and their statistics in the annotated dataset.

Entity Types	
Advice to Others	273
Age	107
Alcohol: Co-ingestion or Amount or Frequency	21
Amount	535
Clinical Impacts	246
Co-ingestion	19
Country of Residence	6
Education Level	17
Ethnicity	8
Gender	70
Household Income	10
IDU: Switch From or Instead of Prescription or In Addition ..	30
Illegal Drug Use	664
Location	193
Marital Status	149
Medical Condition	388
Medicine Intake	1,372
Method of Intake	301
Nonmedical Prescription Drug Use	412
Occupation	49
Relapse	67
Social Impacts	72
Source of Drug	20
Supplements	162
Tobacco Use	19
Transition From Use to Abuse/misuse	13
Transition to IDU	3
Vape Flavor	2
Vape Use	34

Figure 1: Entity types and the number of posts in each entity type.

Datasets	Entity Types	Training Size	Test Size	Entities
Reddit-Impacts	Clinical Impacts,	30k tokens	6k tokens	0.2k tokens
	Social Impacts	1,102 posts	278 posts	318 posts

Table 1: Statistics of Reddit-Impacts dataset, including training and test sizes, the number of entity types and the number of entities in the dataset.

The annotation process of our extensive dataset highlighted the prevalence of readily identifiable concepts such as medicine intake and illegal drug use. However, it also revealed that instances of clinical and social impacts—central to our study—are notably scarce. This scarcity poses significant challenges for research, as these impacts are crucial for understanding the broader consequences of nonmedical substance use on individual health and societal dynamics. To address these challenges and align with our objective of developing more effective public health strategies, we have concentrated our efforts on these two underrepresented entity types, thereby creating the specialized Reddit-Impacts dataset. This focused approach aims to enhance our ability to detect and study these rare but critical impacts in the discourse surrounding substance use.

2.3 Reddit-Impacts Dataset

Index	Span	Token	Entity or not	Label
85-1	13055-13057	In	_	_
85-2	13058-13060	PA	_	_
85-3	13061-13063	at	*	Clinical Impacts
85-4	13064-13065	a	*	Clinical Impacts
85-5	13066-13068	28	*	Clinical Impacts
85-6	13069-13072	day	*	Clinical Impacts
85-7	13073-13078	detox	*	Clinical Impacts
85-8	13078-13079	/	*	Clinical Impacts
85-9	13079-13084	rehab	*	Clinical Impacts
85-10	13085-13089	they	_	_
85-11	13090-13094	used	_	_
85-12	13095-13104	methadone	_	_
85-13	13105-13107	to	_	_
85-14	13108-13111	get	_	_
85-15	13112-13114	me	_	_
85-16	13115-13118	off	_	_
85-17	13119-13121	of	_	_
85-18	13122-13126	bupe	_	_
85-19	13126-13127	.	_	_

Table 2: A sample post "In PA at a 28 day detox / rehab they used methadone to get me off of bupe." with index, spans, tokens and corresponding labels.

From the total of 26,126 posts, only 318 posts (approximately 1.22%) were annotated as having clinical or social impacts. This extremely low occurrence rate underscores the sparsity of relevant data within the larger dataset. Due to the vast size and sparse nature of the original dataset, we opted to randomly select a subset of 1,380 posts for our experiments. We divided the annotated data into 3 sets: 60% for training, 20% for validation, and 20% for testing/evaluation. In summary, Reddit-Impacts comprises 843 posts for training, 259 for validation, and 278 for testing.

This approach not only made the data more manageable but also ensured a focused analysis on the most relevant instances. By narrowing our dataset, we could intensify our efforts on enhancing the detection and classification of these rare but significant entities. This refined dataset formation was pivotal for our experiments and subsequent release of the Reddit-Impacts dataset for the SMM4H 2024 shared task, aiming to provide a resource that is both concentrated and rich in the entities

of interest—clinical impacts and social impacts. The number of instances of our REDDIT-IMPACTS dataset are also shown in Table 1. In addition, table 2 presents an example of posts and their labels.

2.4 Named Entity Recognition

2.4.1 Models

Transformer-based approaches that use large pre-trained language models achieve state-of-the-art F_1-scores for NER tasks when large annotated data available. Due to the sparsity of annotated samples in the REDDIT-IMPACTS dataset, we choose to fine-tune and evaluate two popular transformer-based models as the baseline: BERT [10] and RoBERTa [11]. Building on prior research in few-shot learning, we also report performances for DANN (Data Augmentation with Nearest Neighbor classifier) [12], which demonstrated promising performance in few-shot scenarios.

Given the remarkable success of LLMs in few-shot learning scenarios, we also explore the viability of employing GPT-3.5 for the extraction of named entities in a one-shot setting (by providing one example of input data in the prompt). This evaluation aims to provide insights into the performance of GPT-3.5, further enriching our benchmarking of this dataset.

The following is an outline of the models we used:

1. BERT [10]: A foundational Transformer-based model, widely recognized for its pre-training on a large corpus of text from books and Wikipedia.

2. RoBERTa [11]: Transformer-based model popular for its training on big batches and long sequences.

3. DANN [12]: A few-shot learning method for NER that uses a data augmentation module combined with a nearest neighbor classifier to solve data sparsity problems.

4. GPT-3.5: An advanced iteration of the Generative Pre-trained Transformer series, known for its enhanced language understanding and generation capabilities, trained on a diverse range of internet text.

2.4.2 Evaluation Metrics

We compared the performances of the models based on the micro-averaged F_1-score for clinical impacts and social impacts. We focused our evaluation on these two entity types since that is our class of interest. We report overall entity-level relaxed F_1-score, entity-level strict F_1-score, and token-level F_1-score on these two entity types. For entity-level relaxed F_1-scores, we use SemEval guidelines [2] to calculate partial matches between the predictions and gold-standard annotations.

3 Results and Discussions

3.1 Data and annotation

23% of the posts in the dataset contain words or phrases marked as clinical impact or social impact, with 184 entities annotated as clinical impacts and 67 entities as social impacts.

3.2 Performance on NER Task

Model	Training Size	Entity-level Relaxed F1-Score	Entity-level Strict F1-Score	Token-level F1-Score
BERT	Full training data	0.0	0.0	0.0
RoBERTa	Full training data	0.0	0.0	0.0
DANN	Full training data	**54.36**	**32.62**	**50.79**
GPT-3.5	One-shot	**16.73**	**10.98**	**26.10**

Table 3: Performance on baseline models, including training size we used, entity-level relaxed F_1-score, entity-level strict F_1-score, and token-level F_1-socre.

[2] https://www.davidsbatista.net/blog/2018/05/09/Named_Entity_Evaluation/

Table 3 presents the results of our automatic NER experiments. The table shows the overall relaxed F_1-score, strict F_1-score, and token-level F_1-score on two entity types: clinical impacts and social impacts. The DANN model achieved the highest F_1-score among all the models when the entire training data was used for training. We found that BERT and RoBERTa are unable to identify clinical and social impacts in the dataset despite using the full training data for fine-tuning.

In the few-shot settings, GPT-3.5 tends to perform well on the dataset, even though the evaluation was carried out in a one-shot setting. This demonstrated the significantly higher accuracy of LLMs such as GPT-3.5 in few-shot settings over fine-tuned pre-trained language models (PLMs) for entity extraction in clinical text.

4 Conclusions

Our annotation effort highlighted that information about the clinical and social impacts of substance use are available on Reddit, albeit being sparse. Our experiments demonstrated the difficulty of automatically detecting the sparse clinical and social impact concepts via supervised machine learning, although the DANN model showed promising performance. There is room for improvement in this field, our future efforts will focus on leveraging advancements in large language models like Meta Llama 3 [3] to improve automatic NER performance and evaluate the applicability of our research in real-world settings.

Supplementary

Listing 1: Example prompt we used for GPT-3.5

> You are a medical AI trained to identify and classify tokens into three categories: Clinical Impacts, Social Impacts, and Outside ('O'). 'Clinical Impacts' refer to tokens describing the effects, consequences, or impacts of substance use on individual health or well-being, as defined in UMLS. 'Social Impacts' describe the societal, interpersonal, or community-level effects, also based on UMLS definitions. Any token not falling into these categories should be labeled as 'O'.
> For example, the sentence 'I was a codeine addict.' is tokenized and labeled as follows:
> ['I', 'was', 'a', 'codeine', 'addict', '.'] with labels ['O', 'O', 'O', 'Clinical Impacts', 'Clinical Impacts', 'O'].
> Your task is to predict and return the label for each provided token, ensuring the number of output labels matches the number of input tokens exactly. The output format should be tokens with their labels: ['I-O', 'was-O', 'a-O', 'codeine-Clinical Impacts', 'addict-Clinical Impacts', '.-O'].

References

[1] Laura Lander, Janie Howsare, and Marilyn Byrne. The impact of substance use disorders on families and children: from theory to practice. *Social work in public health*, 28(3-4):194–205, 2013.

[2] Monica Luciana, James M Bjork, Bonnie J Nagel, Deanna M Barch, Raul Gonzalez, Sara Jo Nixon, and Marie T Banich. Adolescent neurocognitive development and impacts of substance use: Overview of the adolescent brain cognitive development (abcd) baseline neurocognition battery. *Developmental cognitive neuroscience*, 32:67–79, 2018.

[3] C Brendan Clark, Cosmas M Zyambo, Ye Li, and Karen L Cropsey. The impact of non-concordant self-report of substance use in clinical trials research. *Addictive behaviors*, 58:74–79, 2016.

[4] Louisa Degenhardt and Wayne Hall. Extent of illicit drug use and dependence, and their contribution to the global burden of disease. *The Lancet*, 379(9810):55–70, 2012.

[3] https://github.com/meta-llama/llama3

[5] Yuting Guo, Swati Rajwal, Sahithi Lakamana, Chia-Chun Chiang, Paul C Menell, Adnan H Shahid, Yi-Chieh Chen, D Pharm, Nikita Chhabra, Wan-Ju Chao, et al. Generalizable natural language processing framework for migraine reporting from social media. *AMIA Summits on Translational Science Proceedings*, 2023:261, 2023.

[6] Abeed Sarker. Social media mining for toxicovigilance of prescription medications: End-to-end pipeline, challenges and future work. *arXiv preprint arXiv:2211.10443*, 2022.

[7] Kayla B Rhidenour, Kate Blackburn, Ashley K Barrett, and Savanna Taylor. Mediating medical marijuana: exploring how veterans discuss their stigmatized substance use on reddit. *Health Communication*, 37(10):1305–1315, 2022.

[8] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

[11] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[12] Yao Ge, Mohammed Ali Al-Garadi, and Abeed Sarker. Data augmentation with nearest neighbor classifier for few-shot named entity recognition. In *MEDINFO 2023—The Future Is Accessible*, pages 690–694. IOS Press, 2024.