Characterizing Information Seeking Events in Health-Related Social Discourse

Omar Sharif[1], Madhusudan Basak[1], Tanzia Parvin[2], Ava Scharfstein[1], Alphonso Bradham[1], Jacob T. Borodovsky[3, 4], Sarah E. Lord[3, 4, 5], Sarah M. Preum[1, 3, 4]

[1]Department of Computer Science, Dartmouth College
[2]Department of Computer Science and Engineering, CUET, Bangladesh
[3]Center for Technology and Behavioral Health, Geisel School of Medicine, Dartmouth College
[4]Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College
[5]Department of Psychiatry, Dartmouth Health
{omar.sharif.gr, sarah.masud.preum}@dartmouth.edu

Abstract

Social media sites have become a popular platform for individuals to seek and share health information. Despite the progress in natural language processing for social media mining, a gap remains in analyzing health-related texts on social discourse in the context of events. Event-driven analysis can offer insights into different facets of healthcare at an individual and collective level, including treatment options, misconceptions, knowledge gaps, etc. This paper presents a paradigm to characterize health-related information-seeking in social discourse through the lens of events. Events here are board categories defined with domain experts that capture the trajectory of the treatment/medication. To illustrate the value of this approach, we analyze Reddit posts regarding medications for OPIOID USE DISORDER (OUD), a critical global health concern. To the best of our knowledge, this is the first attempt to define event categories for characterizing information-seeking in OUD social discourse. Guided by domain experts, we develop *TREAT-ISE*, a novel multilabel treatment information-seeking event dataset to analyze online discourse on an event-based framework. This dataset contains Reddit posts on information-seeking events related to recovery from OUD, where each post is annotated based on the type of events. We also establish a strong performance benchmark (77.4% F1 score) for the task by employing several machine learning and deep learning classifiers. Finally, we thoroughly investigate the performance and errors of ChatGPT on this task, providing valuable insights into the LLM's capabilities and ongoing characterization efforts.

Introduction

About 70% of people in the USA rely on social media to connect with peers and share information to navigate aspects of their lives (**?**), such as finance, professional development, and health. Approximately 72% of adult internet users engage in online searches to find information about various health issues, either for themselves or on behalf of others. Individuals from diverse socio-demographic backgrounds and medical conditions often resort to these platforms for sharing and seeking information during different phases of their healthcare journey (**?**). As a result, online platforms offer a unique opportunity to surface and contextualize information needs, knowledge gaps, treatment perceptions, conflicting information, and misconceptions as illustrated in prior works (**????**).

Event analysis is a well-studied NLP problem that can be leveraged to get insights into online health discourse (**??**). Analyzing posts from social support groups through the lens of events can give us a nuanced understanding of health information needs. For example, consider the following post seeking information on post-operative pain relief.

> I just had knee surgery about a week ago, and my **pain meds** (**Gabapentin**) are not cutting it! I am having **sleep troubles** and getting pretty **anxious** about recovery. I am considering **Kratom**. Any suggestions about **how much I can take per day** and **how often**?

This post mentions multiple key events, e.g., taking medication (*Gabapentin*), and experiencing psychophysical effects (*pain* or *anxiety*). Such event analysis based on a large number of samples can reveal insights into different aspects of treatment at both individual and collective levels (e.g., *how many people report a new or rare side effect during pain management*), the perceived value of treatment (e.g., *ineffective pain medication*), self-treatment strategies (e.g., *self-dosing Kratom*), knowledge gaps and concern (*rare or new side effects of treatment*), and misconceptions (e.g., *self-dosing Kratom is safe for pain relief*). However, a significant gap still exists in performing event-driven analysis on health discourse in online communities, including social media.

To showcase the significance of event-driven analysis, we explore online discussions regarding recovery from opioid use disorder (OUD), a critical concern with substantial societal impact. OUD remains a leading cause of mortality in the US, incurring a massive economic toll, estimated at 1.02 trillion dollars annually (**?**). Existing challenges, including stigma around addiction, limited healthcare access, and distrust of traditional systems, drive individuals towards seeking recovery support within social groups (**?**). Medications for opioid use disorder (MOUD) offer a vital treatment avenue, capable of saving lives (**?**). Pseudo-anonymous platforms like Reddit, known for its wide US user base and emphasis on anonymity, provide a unique lens into MOUD-based recovery insights. Reddit's reach and focus on sensitive topics, such as mental health and substance use disorder, position it as a significant source for rich content (**???**).

Event analysis on social media text confronts distinct

hurdles. Defining and standardizing event types pose a challenge, especially considering the influence of domain-specific factors that determine event relevance*clinical* events differing from *career* events, for instance. Moreover, equivalent events might be expressed in vastly dissimilar colloquial terms. Remarkably, there exists no dataset for delving into event analysis within online discourses.

This paper introduces an event-based framework for online discourse analysis. We study Reddit posts on MOUD, a critical topic amid the opioid crisis. Collaborating with experts, we define information-seeking events, create annotation guidelines, and curate a unique labeled dataset. Using the event schema and info-seeking posts, we explore information quality systematically. Comprehensive insight into treatment needs relies on classifying data into specific events. We frame identifying core events in posts as a multi-label, multi-class classification challenge. We assess the dataset with advanced text classifiers including large language models. Our major contributions are as follows.

- **Resource:** Based on guidance from domain experts, we propose a treatment information-seeking event (ISE) schema that can help to understand the OUD treatment trajectory. Leveraging this schema, we develop *TREAT-ISE*, a multilabel dataset comprising human-annotated samples. The novel dataset, annotation guidelines, and associated code will accelerate further research in online health discourse analysis[1].
- **Social:** We focus on a highly vulnerable population, i.e., individuals considering or undergoing OUD recovery, which has received little attention in previous work, and characterize their self-reported MOUD treatment information needs. The dataset and other outcomes can complement traditional electronic health records and survey data and capture the real-world complexity of recovery.
- **Benchmarking:** We investigate the performance of ten off-the-shelf machine learning and deep learning models for this task. Furthermore, we thoroughly assess the effectiveness of ChatGPT, thereby uncovering the potential scope of ChatGPT and state-of-the-art text classifiers for such complex, knowledge-intensive discourse analysis.

Related Work

Social Media and Substance Use Disorder

Social media platforms offer individuals opportunities to share the different events of their lived experiences, such as addiction, logistical barriers, treatment strategy, the experience of psychophysical effects, and more (**?**). Prior studies have utilized these data to understand different types of substance usage, including cannabis, alcohol, opioids, and others (**??**). **?** (**?**) tried to uncover alternative treatment options for OUD by analyzing the opioid discourse on Reddit. **?** (**?**) presented a framework for extracting keywords and applied it to extract insights about OUD recovery from Reddit. Another related study by **?** (**?**) investigated how much online social community can support individuals undergoing opioid usage. Our work differs from existing studies both in task formulation and scope. We focus on events as the primary analytical lens, encompassing diverse categories to study OUD discourse.

Event Analysis for Health-related Discourse

Event analysis includes two main subtasks: *event detection:* identifying trigger terms and event type, and *argument extraction:* extracting event arguments from texts and assigning roles to arguments based on event type (**?**). Recent approaches have adopted a text-generation paradigm, leveraging large language models to prompt the extraction of event types, triggers, and arguments (**??**). Few works attempted to detect events without extracting triggers (**?**). However, these models often exhibit suboptimal performance when dealing with event analysis tasks heavily reliant on domain-specific knowledge (**?**). The limited availability of training data and the complexity of domain-specific terms contribute to this issue.

Research on event analysis from social media discourse regarding health is limited. **?** (**?**) attempted to develop a disease progression timeline by analyzing patient-authored texts in social support groups. They explored the connections between medical events and users' engagement within these groups. In a similar work, **?** (**?**) investigated the behavioral trajectory of participants by analyzing cancer-related events in online medical discourse. In subsequent work, **?** (**?**) created a temporal tagger to extract cancer-related event dates to explore treatment trajectories.

This work focuses on event detection rather than trigger or argument extraction since it is one of the first attempts to understand OUD discourse in the context of events. Here, we formulate the event detection task as a multi-label, multi-class classification problem (**?**). We curate a dataset of information-seeking events by actively involving domain experts in the process. Moreover, we analyze ChatGPT's performance and errors, presenting valuable insights into the model's capabilities and contributing to ongoing efforts to understand its characteristics.

Defining Information-Seeking Events

Our overarching goal is to characterize information-seeking events (ISE) from online discourse. These events are self-reported by individuals considering or undergoing medications for OUD (MOUD) treatment. MOUD includes Buprenorphine (e.g., Suboxone, Subutex, Sublocade), Methadone, and Naltrexone (**?**). We collaborate with five domain experts to define the events that best characterize the treatment information-seeking events from different stages of the treatment journey. Our collaborators are well-versed and internationally acclaimed scholars in substance use disorder, spanning various areas such as epidemiology, public health policy, mental health, addiction psychiatry, addiction medicine, and biomedical data science. They review the collected samples and offer valuable insights into various facets of opioid recovery. Based on the guidance from the domain experts, we identify five coarse categories of events for treatment information needs. These ISE categories are: *Accessing MOUD*, *Taking MOUD*, *Experiencing*

[1]All the resources are available at https://github.com/omar-sharif03/AAAI-2024

Psychophysical Effects, *Relapse or co-occurring substance usage*, and *Tapering MOUD*. All of these events are prevalent in recovery using MOUD.

- **Accessing MOUD (AM):** Information-seeking events related to accessing MOUD, such as concerns about insurance, pharmacy, providers, etc. Analyzing samples from this event can help to determine the common barriers people encounter during recovery using MOUD that affect treatment induction, adherence and retention.
- **Taking MOUD (TM):** Information-seeking events related to MOUD regimen details, e.g., questions about timing, dosage, frequency of taking a MOUD, concerns about splitting and missing a dose. This class can surface potential misconceptions and concerns about MOUD administration that negatively impact treatment adherence.
- **Experiencing Psychophysical Effects during Recovery (EP):** Information-seeking events related to concern about potential physical and/or psychological effects during recovery. This event class covers both experienced and anticipated psychophysical effects. It can surface rare and new adverse effects of MOUD as well as prevalent psychophysical effects of MOUD, their severity and potential impact of treatment adherence.
- **Relapse (RL) or co-occurring substance usage:** This class includes events that talk about relapsing or using other substances during recovery. Such substance use can be attributed to recreational purposes or for self-medication (e.g., marijuana for sleep). We follow NIDA's[2] list of commonly used drugs to identify what counts as a substance. Samples of this event class can help unearth specific information individuals seek concerning recreational and medical usage of substances.
- **Tapering MOUD (TP):** Information-seeking events related to reducing the dose or frequency of MOUD and eventually quitting MOUD. Although the current standard of care recommends consulting healthcare providers for tapering MOUD, individuals often resort to self-managed tapering strategies. Analyzing events from this class can inform addiction researchers and clinicians about the context of self-managed tapering strategies (e.g., why and when people self-taper) and their effectiveness (what works for whom).
- **Others (Oth):** Information-seeking events related to other issues.

In this work, we focus on information-seeking events from posts on social media. It should be noted that we can also analyze relevant information-providing or sharing events (i.e., comments or replies to the original post) through this lens of events. This will help us measure the availability and quality of shared information in online discourse more systematically, e.g., self-management strategies for tapering, high-dose of MOUD suggested by peers or common misinformation regarding relapse during recovery using MOUD. Such systematic analysis can potentially uncover actionable insights to improve treatment adherence and outcomes.

[2]https://tinyurl.com/4ckwz453

Dataset Collection and Annotation Strategies

Data collection

We chose Reddit as the data source due to its anonymity policy and rich content on MOUD (**?**). We selected r/Suboxone as our primary data source as (i) it has both the highest number of members and the number of peer interactions (e.g., number of posts, comments) among the subreddits specific to different options of MOUD; and (ii) it is strictly moderated where any irrelevant posts are removed by moderators (e.g., drug soliciting posts). So, this subreddit offers a unique chance to understand users' information needs related to a MOUD authentically. We scraped all the posts between January 2018 (as minimal interaction was observed in this subreddit prior to 2018) and August 2022 (study start time). We collected a total of 25,044 posts using the PRAW and PushShift APIs (**?**). The collected data includes titles, posts, comments, likes, upvotes, and unique post IDs while strictly adhering to ethical considerations by not collecting/storing information that violates ethical concerns. After removing irrelevant posts (e.g., polls, link-only posts), we ended up with a corpus of 15,253 relevant posts. Among these posts, we annotate 5083 randomly selected posts.

Data Annotation

Feasibility of Crowd-sourced Annotation: Annotating the type of treatment information-seeking events is a challenging task that demands significant effort. Initially, we employed the widely-used approach of annotating through crowd-workers on Amazon Mechanical Turk (**?**). We selected a pool of Master qualified workers (mTurkers with high approval ratings), provided them with explicit annotation guidelines, and conducted a trial run on 300 samples. However, we encountered poor annotation quality and low inter-annotator agreement (only 40.5%). This is because this annotation task requires a good understanding of domain knowledge and annotators need interactive, progressive training sessions to ensure they understand the nuances of different types of events. Our trial run indicates a lack of suitability of crowd-workers for such a challenging inference task. Therefore, we decided to perform in-house annotation with students and experts.

Annotation process: To complete the annotation, we form a diverse group of 9 annotators: 3 undergraduates and 6 graduate students. Initially, we provided them with background knowledge on MOUD and suboxone through multiple sessions led by experts. To achieve quality annotation, our primary focus was to confirm that the annotators understand what are the ISEs for OUD recovery and how a user can seek information for multiple event types in a post (details added in the appendix[3]). Each annotator was trained for four weeks through trial annotation tasks before they started actual annotation to ensure annotators were well-versed with ISE classes and eliminate the uncertainties about annotation guidelines. Each sample was reviewed by at least two different annotators for the annotation. Table 1 demonstrates sample posts with associated labels.

[3]https://tinyurl.com/yyrn7ywn

Title	Post	Events
Looking for suboxone guidance?	I take 1-2mg subs per day which is a decrease from the original dose of 8mg. Just looking for a plan of action in which to stick with to eventually get off completely.	Taking MOUD (TM), Tapering (TP)
Which Kratom strain helps with Bupe withdrawal	When I run out of my Suboxone prematurely, I like to keep Kratom on hand for my extremely low energy and excessive yawning.	Relapse (RL), Psychophysical effects (EP)

Table 1: Sample data excerpts with titles, posts, and labels (shortened and paraphrased as per IRB guideline).

Inter-annotator Agreement: We compute the inter-annotator agreement in terms of Cohen's κ-score (**?**). Table 2 shows the κ-score for each class where the AM class achieves the highest agreement score of 0.86, and the *other* class gets the lowest (0.68). The mean κ-score of 0.76 indicates substantial agreement between the annotators. The presence of domain-specific drug names, lengthy text samples with shorthand, slang, and misspellings posed challenges during annotation. Table 2 exhibits the initial agreement score between two annotators. It is important to mention that a domain expert reviewed all the samples after labeling by two annotators to ensure the data quality. Subsequently, resolved any confusion or annotation disagreement and rectified the labels. This domain expert is a study team member who is disjointed from the set of recruited annotators. This complies with the recommended best practice for qualitative health data annotation/coding (**?**). Thus we develop **TREAT-ISE** a MOUD treatment information-seeking event dataset comprised of 5083 multilabel samples.

	AM	TM	TP	EP	RL	Oth	Mean
κ-score	0.86	0.72	0.82	0.74	0.75	0.68	0.76

Table 2: Classwise κ-score of TREAT-ISE.

Class	#Samples	#Words	#Unique words	#Avg. words/sample
AM	873	108k	7665	124.16
TM	1637	199k	10477	122.07
TP	1424	215k	11087	151.40
EP	1837	271k	13395	147.75
RL	1420	202k	10776	142.33
Oth	473	48k	6159	102.62

Table 3: Summary of different classes of the TREAT-ISE dataset.

Table 3 presents the statistics and lexical summary of TREAT-ISE. The dataset is imbalanced, with EP having the highest number of samples. Among the classes, EP stands out with the most words ($\approx$271k) and unique words ($\approx$ 13k), while the AM class has the lowest counts ($\approx$108k, $\approx$7.6k). TREAT-ISE stands apart from other domain-specific datasets by presenting a unique multilabel classification challenge with significantly longer average sample lengths (ranging from 122 to 151). The average sample length for similar multilabel classification tasks is less than 50 (**??**). In the ablation studies, we present a few insights into how large language models handle these domain-specific long texts.

Methodology

We present a comprehensive benchmark evaluation of the TREAT-ISE dataset encompassing various methodologies, including off-the-shelf non-transformer, Transformer-based, and Large language models such as ChatGPT. These methods represent standard approaches for multilabel classification and provide a diverse range of technical implementations for thoroughness. The details of each method are described in the subsequent paragraphs.

- **Non-transformer models:** In the baseline evaluation, we explore the performance with two machine learning models: Logistic Regression (LR) (**?**) and Naive Bayes with Support Vector Machine (NBSVM) (**?**). For deep learning approaches, we investigate two variants: one utilizes pretrained FastText (**?**) embeddings with a feed-forward network, and the other employs a Bidirectional Gated Recurrent Unit (BiGRU). In BiGRU, embedding features are propagated to a GRU layer with 80 hidden units. The output from the last hidden layer is passed to global average pooling and max-pooling layers. Subsequently, the outputs of the pooling layers are concatenated and passed for classification.
- **Transfomer-based models:** In recent years, transfomer-based (**?**) models have achieved state-of-the-art performance on various NLP tasks. We employ six transformer-based pre-trained models to benchmark the multilabel ISE classification task. These include Bidirectional Encoder Representations from Transformers (BERT) (**?**), a distilled variant of BERT (DistilBERT) (**?**), robust BERT architectures with more training data RoBERTa (**?**), ELECTRA (**?**), a model with generalized autoregressive pertaining (XLNet) (**?**) and MPNet (**?**). All the models are sourced from the Huggingface library. Subsequently, fine-tuned on our dataset for 10 epochs with a learning rate $2e^{-5}$ and batch size 16. The intermediate model demonstrating the best validation set performance is saved for the test set prediction.
- **ChatGPT:** Several recent studies have demonstrated that large language models like ChatGPT can surpass humans in various classification and annotation tasks (**??**). So we explore the scope of ChatGPT (**?**) to classify ISE in our annotated dataset. To comprehensively assess its capabilities, we explore three distinct settings: zero-shot (ZS), few-shot (FS), and chain-of-thought (CoT) (**?**) prompting. The chain-of-thought approach gives the

	AM			TM			TP			EP			RL			
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**WF1**
Non-transformer Baselines																
LR	0.63	0.68	0.65	0.61	0.64	0.63	0.74	0.64	0.68	0.49	0.66	0.56	0.71	0.57	0.63	0.593
NBSVM	0.73	0.57	0.64	0.58	0.73	0.65	0.72	0.78	0.75	0.44	0.75	0.56	0.61	0.62	0.62	0.602
FastText	0.64	0.73	0.68	0.61	0.70	0.65	0.69	0.84	0.75	0.50	0.78	0.61	0.63	0.64	0.63	0.624
BiGRU	0.72	0.69	0.70	0.73	0.68	0.70	0.80	0.84	0.82	0.61	0.61	0.61	0.80	0.72	0.76	0.702
Transformer Baselines																
BERT	0.85	0.74	0.79	**0.83**	0.64	0.72	0.84	0.84	0.84	**0.69**	0.59	0.64	0.88	0.76	0.81	0.733
RoBERTa	0.82	0.82	0.82	0.75	0.80	**0.77**	0.80	0.89	0.84	0.63	0.74	**0.68**	0.88	0.75	0.81	0.757
Distil-BERT	0.80	0.69	0.74	0.81	0.62	0.70	0.84	0.78	0.81	0.68	0.57	0.62	0.84	0.76	0.80	0.711
ELECTRA	0.77	0.79	0.78	0.80	0.67	0.73	**0.87**	0.84	0.85	0.65	0.64	0.65	0.80	**0.88**	0.84	0.748
XLNet	0.84	**0.82**	**0.83**	0.79	0.72	0.75	0.85	0.84	**0.85**	0.59	0.78	0.67	**0.88**	0.85	**0.86**	**0.774**
MPNet	0.79	0.81	0.80	0.80	0.71	0.75	0.81	0.85	0.83	0.68	0.66	0.67	0.78	0.82	0.80	0.751
ChatGPT Baselines																
ChatGPT (ZS-S)	**1.0**	0.26	0.41	0.74	0.30	0.43	0.67	0.42	0.52	0.62	0.10	0.18	0.62	0.81	0.70	0.433
ChatGPT (ZS-L)	0.78	0.61	0.69	0.68	0.63	0.65	0.69	0.67	0.68	0.70	0.29	0.41	0.77	0.53	0.63	0.581
ChatGPT (FS-S)	0.48	0.79	0.60	0.47	**0.92**	0.62	0.45	**0.96**	0.61	0.44	0.83	0.57	0.65	0.69	0.67	0.609
ChatGPT (FS-L)	0.51	0.78	0.62	0.52	0.87	0.65	0.50	0.86	0.63	0.40	**0.92**	0.56	0.66	0.72	0.69	0.620
ChatGPT (CoT)	0.62	0.78	0.69	0.49	0.87	0.62	0.55	0.89	0.68	0.49	0.76	0.60	0.74	0.56	0.64	0.631

Table 4: Classwise performance for treatment information seeking event detection. WF1 indicates the weighted F1 score based on all six classes. The shorthand indicates ZS-S, ZS-L: Zero-shot (Short, Long), FS-S, FS-L: Few-shot (Short, Long), and CoT: Chain-of-Thought prompting. Due to space constraints, the models' performance in *other* class is not included.

model more reasoning about this domain-specific task (**?**). We thoroughly explored various versions of prompts and refined those that showed encouraging outcomes. We select the optimal prompt through an iterative process of trial and error guided by the empirical observations of the model's output. Our approach involves two prompt templates for conducting the experiments: *'Short'* and *'Long'*. The *'Short'* template offers minimal details concerning the ISE classes, while the *'Long'* variant provides the model with a detailed definition of the classes. We adopt both *'Short'* and *'Long'* templates in the ZS and FS experiments. However, for the chain-of-thought approach, we use a modified version of long prompts, including the reasoning for the examples. We set the temperature value to 0.0 across all experiments to ensure the deterministic behavior of the model.

The test set **excludes** all the samples used to identify the best prompts and in-context examples used in the few-shot and chain-of-thought prompts. This ensures unbiased evaluation and prevents the risk of potential data leakage. Due to space constraints, we could not share example prompts in the main paper. However, they are readily accessible through the appendix[4].

TREAT-ISE: Benchmark Evaluation

In this section, we outline the evaluation settings and present the results. We perform comprehensive ablation studies to understand the performance of advanced LLMs, like ChatGPT, on domain-specific, complex text classification task.

Experimental and Evaluation Setup: All the experiments were conducted on a GPU-accelerated Google Colab platform. Machine learning and deep learning models were trained with ktrain (**?**), while all the transformer models were implemented from Huggingface. Finally, we investigate the performance of the ChatGPT model via API (version *gpt-3.5-turbo-0613*) calls.

TREAT-ISE is partitioned into three mutually exclusive sets: train (80%), validation (10%), and test (10%). We leverage various statistical measures (precision (P), recall (R), F1-score) to asses and understand the model's performance across different classes. The validation set is utilized to tune the model hyperparameters across various experiments. The weighted F1-score (WF1) on the test set is used to compare and determine the superiority of the models.

Results

Table 4 presents the classwise performance of all the models on the test set of the TREAT-ISE dataset. BiGRU achieved the highest WF1 of 0.702 among the non-transformer baselines. XLNet outperformed all other models with a maximal WF1 score of 0.774. It excelled particularly well in AM, TP, and RL classes, with scores of 0.83, 0.85, and 0.86, respectively. RoBERTa attained the highest WF1 of 0.757 of the BERT variants. All models encountered challenges in identifying samples from *taking medication* (TM) and *experiencing psychophysical effects* (EP) events. Surprisingly, all the ChatGPT variants underperformed compared to other baselines. We conducted a detailed ablation study to get more insights into this. The CoT prompt acquired the highest WF1 of 0.631 and outperformed all other prompting techniques. In contrast to the other baselines, where there is a balance between classwise precision and recall, all the ChatGPT variants (except ZS-S) showed much higher recall than precision. This indicates that ChatGPT tends to overpredict the classes. Overall, the results indicate that identifying treatment information-seeking events is difficult, requires domain knowledge, and has significant room for improvement.

[4]https://tinyurl.com/ycyj3v4a

Statistical significance: We conduct statistical significance testing using the **?** (**?**) test to see if the best-performing model (i.e., XLNet) outperforms other models in a statistically meaningful way. Since the results from each classifier are nominal data (i.e., classes of events) and it is difficult to train multiple copies of a model, this test is a suitable approach. We conducted a pair-wise comparison between XLNet and all the other models for each event class. The classwise *P-value* indicates XLNet is significantly better than all other models in three of the five classes, namely, TM (P<0.001), EP (P=0.008), and TP (P=0.003). The performance of XLNet is not statistically significant for the remaining two classes. Specifically for AM (P=0.657) and RL (P=0.446), XLNet's performance is comparable to RoBERTa and FastText, respectively.

Ablation Studies with ChatGPT and XLNet

Although recent studies illustrate that ChatGPT can outperform humans in knowledge-intensive tasks (**?**), results (Table 4) demonstrate that ChatGPT exhibits suboptimal performance in classifying treatment information-seeking events. This is particularly noticeable for samples that require significant domain knowledge to distinguish between events. This motivates us to conduct a deeper investigation into the scope of ChatGPT for such event analysis. We also aim to uncover whether the errors of the transformer models are echoed in ChatGPT or they are distinct. So, in this section, we perform a thorough side-by-side analysis between the best-performing model in our task, XLNet, and the top-performing prompt setting of the ChatGPT model (i.e., chain-of-thought). The findings are as follows.

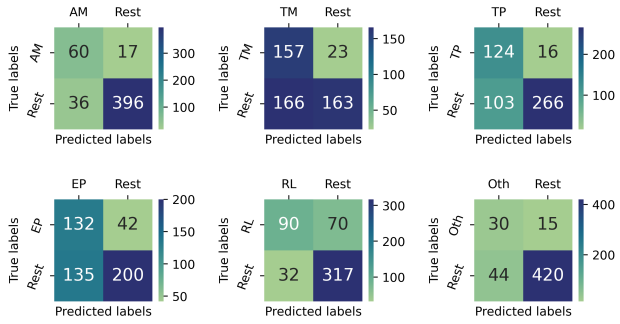

Figure 1: Confusion matrices of each category for ChatGPT with the chain-of-thought (CoT) approach. The *Rest* class indicates predictions on all other classes.

- **ChatGPT tends to overpredict more:** After qualitative and quantitative analysis, we found that ChatGPT often fails to understand the context holistically and overpredicts. Consider the following example, which is about *relapse* (RL) and *tapering* (TP). Although ChatGPT predicted these labels correctly because of the mention of side effects and dosage information, it erroneously added TM and EP labels.

  ...I started by quitting kratom completely and taking 2mg of suboxone, I experienced no **withdrawals** during the switch but also no high. Today I'm down to **1.5mg of suboxone**, and I'm so happy! Planning to go down to 1.25mg pretty soon too.

Figure 1 illustrates the confusion matrices for the ChatGPT chain-of-thought approach. Table 5 presents the classwise overprediction ratio (#false positive / #predicted positives) for both ChatGPT and XLNet. Surprisingly, the average overprediction ratio for ChatGPT is 45%. That means almost half of the time, it incorrectly predicts that samples contain information-seeking events. ChatGPT exhibits higher error in the TM and EP classes, with 166 (out of 323) and 135 (out of 267) mispredictions, respectively. In contrast, XLNet exhibits a drastically lower overprediction ratio for all categories except in the EP class.

	AM	TM	TP	EP	RL	Oth
CG	36/96 0.375	166/323 0.513	103/227 0.453	135/267 0.505	32/122 **0.262**	44/74 0.594
XL	12/75 0.160	35/165 0.212	21/139 0.151	92/227 0.405	19/155 **0.122**	9/33 0.27

Table 5: Classwise overprediction ratio (#false positive / #predicted positives) of ChatGPT (CG) with CoT prompts and the XLNet (XL) model.

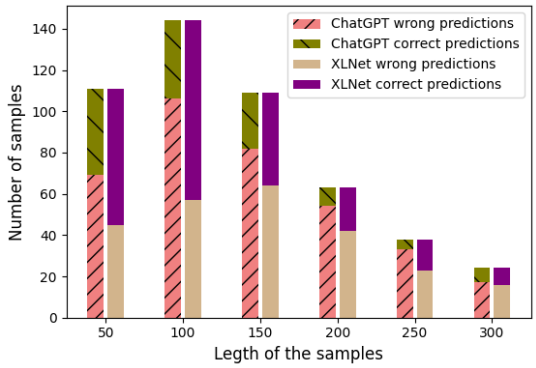

Figure 2: Correlation between sample length and frequency of correct/wrong predictions: as the length of samples (measured in words) increases, the frequencies of accurate predictions decrease for both models.

- **ChatGPT struggles more on long samples:** For analysis, we compute the frequency of correct and wrong predictions on different length ranges for both ChatGPT (CoT) and XLNet. Figure 2 illustrates the correlation, indicating that the frequency of accurate prediction is higher among the shorter samples and decreases as sample length increases. On average, the samples where ChatGPT made errors had a length of 128.04, whereas, for XLNet, this value is 140.21. This analysis suggests both models encounter difficulties in understanding information-seeking events with long-range context. However, XLNet shows slightly more robustness than ChatGPT.

- **ChatGPT misclassifies events more:** To obtain the confusion mapping, we calculate the frequency of incorrect predictions for each event in relation to other events as presented in Table 6. Analyzing the results, it becomes evident that ChatGPT faces difficulty in distinguishing advice events associated with TM, EP, and RL classes, often misclassifying them as *other* class. The model made the highest (92) number of errors on the RL event class and, most of the time considered it as either the TM or EP event class. Interestingly, XLNet often misclassified TM as EP (10) class.

		AM	**TM**	**TP**	**EP**	**RL**	**Oth**	Total
AM	CG	-	**7**	2	3	2	5	19
	XL	-	1	1	3	2	3	10
TM	CG	1	-	2	2	2	**10**	17
	XL	0	-	6	10	5	3	24
TP	CG	2	1	-	2	1	**8**	14
	XL	0	5	-	6	1	0	12
EP	CG	1	6	5	-	3	**20**	35
	XL	1	1	1	-	2	0	5
RL	CG	6	29	15	**31**	-	11	92
	XL	0	5	1	6	-	1	13
Oth	CG	4	**10**	1	4	2	-	21
	XL	7	3	1	1	0	-	12

Table 6: Confusion mapping of ChatGPT (CG) with chain-of-thought approach and XLNet (XL) model. Each cell indicates how many times an event (in row) confuses with another event indicated in the column.

The results suggest that ChatGPT is biased toward predicting TM and EP classes. After qualitative observation, we notice that ChatGPT often mislabels samples as TM when dosage information is provided *(e.g., 2mg kratom, 1.5 mg bupe)*, even though these instances do not seek treatment information. Similarly, the model frequently mislabels posts mentioning psychophysical effects (e.g., withdrawals, sleep) as EP, despite these not being information-seeking events. Surprisingly, on 54 occasions, the model identified posts that were seeking treatment information but failed to predict appropriate event classes and mislabeled them as the *other* class. This mislabeling can be attributed to the model's poor understanding of the domain-specific nuances.

Conclusion and Future Work

In this paper, we address a critical social concern by investigating the information needs of individuals who are considering or undergoing recovery from opioid use disorder. On the guidance of experts, we develop a multilabel, multiclass dataset (*TREAT-ISE*) aiming to characterize OUD treatment information-seeking events. This dataset introduces a new resource to the field, enabling us to study MOUD treatment for recovery through the lens of *events*. The event schema we defined can be valuable to surface clinical insights such as knowledge gaps about treatment, tapering strategies, potential misconceptions, and beyond. Moreover, our data collection process, event-centric schema design, and data annotation strategy can be replicated to develop similar resources for other domains. Finally, we benchmark the dataset with a wide range of NLP models and demonstrate the potential challenges of the task with thorough ablation studies.

There are several scopes for potential improvement. Due to costly and time-consuming annotation, we had to limit the dataset size to 5083 samples. We will explore the possibility of minimal supervision to augment the dataset size by leveraging our annotation protocol and additional available data (over 10K samples). Other research can explore how treatment information-seeking events vary in other online communities and subreddits. In addition, investigating how other large models (e.g., GPT-4, LLaMA) perform on this task can provide us with valuable insights.

Ethical Considerations

This research was approved by the Institutional Review Board (IRB) of the author's institution.

User Privacy: All the data samples were collected and annotated in a manner consistent with the terms and conditions of the respective data source. We do not collect or share any personal information (e.g., age, location, gender, identity) that violates the user's privacy.

Biases: Any biases found in the dataset and model are unintentional. Experts and a set of diverse groups of annotators labeled the data following a comprehensive annotation guideline and all annotations were reviewed to address any potential annotation biases. Our data collection exclusively focused on one subreddit (r/suboxone), possibly leading to a bias towards the r/suboxone community. The developed models can only be used to identify events that we discussed in the paper. So the chance of using these models for malicious reasons is very minimal.

Intended Use: We intend to make our dataset accessible per Reddit policies to encourage further research on online health discourse as well research on MOUD.

Acknowledgement

The preparation of this article was partially supported by P30 Center of Excellence grant from the National Institute on Drug Abuse (NIDA) P30DA029926 (PI: Lisa A. Marsch).

Eveniet animi voluptatibus nobis eum rerum dignissimos maiores ad hic dolorum, asperiores quibusdam dolore est velit nesciunt corrupti similique labore sit harum ipsa, inventore voluptatibus explicabo ex optio reiciendis laboriosam vitae sit molestias, tempore optio voluptas reprehenderit odio tenetur nobis quae exercitationem expedita, aliquam alias corrupti?Beatae est necessitatibus nam nisi, facere error tempore harum consequatur labore quaerat officia pariatur sint?Quaerat illum itaque necessitatibus ipsum ullam modi laudantium eos, praesentium mollitia quod qui hic totam necessitatibus, ab non ipsum impedit, recusandae repudiandae natus omnis veniam soluta unde corporis necessitatibus suscipit voluptas, quo est voluptas cum ullam eius cumque alias?Incidunt tempore voluptatum consequuntur, necessitatibus ea consequatur dolor, odio ad molestias, sapiente sus-

cipit reiciendis iusto at modi quo velit?Numquam delectus in ut dicta labore distinctio et, deleniti animi provident laborum a dolores vero distinctio