context. However, XLNet shows slightly more robustness than ChatGPT.

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