# Decoding the Narratives: Analyzing Personal Drug Experiences Shared on Reddit

**Layla Bouzoubaa, Elham Aghakhani, Max Song, Minh Trinh, Rezvaneh Rezapour**
Drexel University
{lb3338, ea664, ms5526, mqt32, shadi.rezapour}@drexel.edu

## Abstract

Online communities such as drug-related subreddits serve as safe spaces for people who use drugs (PWUD), fostering discussions on substance use experiences, harm reduction, and addiction recovery. Users' shared narratives on these forums provide insights into the likelihood of developing a substance use disorder (SUD) and recovery potential. Our study aims to develop a multi-level, multi-label classification model to analyze online user-generated texts about substance use experiences. For this purpose, we first introduce a novel taxonomy to assess the nature of posts, including their intended connections (Inquisition or Disclosure), subjects (e.g., Recovery, Dependency), and specific objectives (e.g., Relapse, Quality, Safety). Using various multi-label classification algorithms on a set of annotated data, we show that GPT-4, when prompted with instructions, definitions, and examples, outperformed all other models. We apply this model to label an additional 1,000 posts and analyze the categories of linguistic expression used within posts in each class. Our analysis shows that topics such as Safety, Combination of Substances, and Mental Health see more disclosure, while discussions about physiological Effects focus on harm reduction. Our work enriches the understanding of PWUD's experiences and informs the broader knowledge base on SUD and drug use.

## 1 Introduction

**Warning:** This paper includes language and content that may be offensive or triggering.

For people who use drugs (PWUD), social platforms like Reddit serve as invaluable spaces for open discussion and community support. Such platforms enable PWUD to engage in conversations and share experiences, facilitated by the anonymity and community solidarity that Reddit provides (Figure 1). Despite Reddit's ability to foster connection through shared experiences (Bouzoubaa et al., 2023; Choudhury and De, 2014), the perspectives

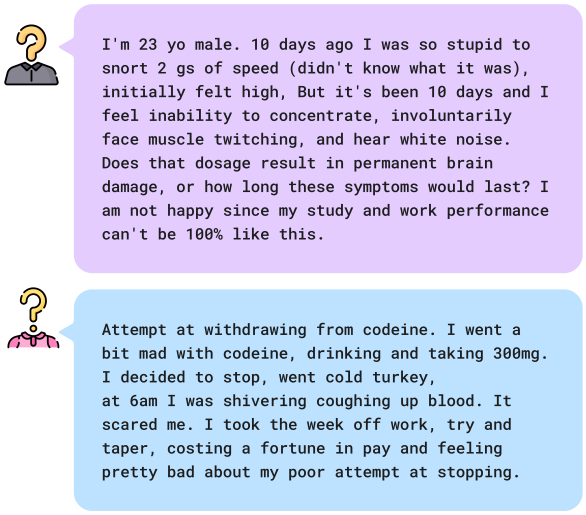



Figure 1: Examples of user-generated posts on drug-related subreddits

of PWUD are frequently marginalized in important decision-making processes, particularly for those who depend exclusively on online communities due to the significant stigma associated with seeking help from traditional services (Volkow et al., 2021).

Analyzing how PWUD communicate about their experiences on online platforms offers important insights into their narratives, highlighting their information-seeking behaviors and the diverse needs specific to this population (Valdez and Patterson, 2022), revealing the real-life challenges and perspectives of individuals dealing with Substance Use Disorders (SUD) (Brown et al., 2019; Bouzoubaa and Rezapour, 2024). Existing literature in this domain primarily focuses on classifying mental health experiences like depression (Rijen et al., 2019) or drug-related events like abuse (Al-Garadi et al., 2021) or overdose risk (Garg et al., 2021). Personal lived experiences of PWUD remain largely unexplored, highlighting the need for effective methodologies to understand and analyze diverse drug experiences shared online.

Our study aims to bridge this gap by exploring a range of substance use experiences shared online,

from recovery to general discussions, through the development of a taxonomy and model for classifying these narratives on Reddit. Insights from this study will not only expand our understanding of how PWUD navigate information seeking and support mechanisms, thereby laying the groundwork for harm reduction strategies and more nuanced, effective interventions, but also help in identifying and understanding the diverse experiences and characteristics of individuals most impacted by drug misuse (Strapparava and Mihalcea, 2017; Yang et al., 2023).

To understand the experiences of PWUD, we first developed a fine-grained taxonomy of online drug experiences comprising three levels: Connection (which highlights the information-seeking and sharing behaviors of users), Subject (lived SUD-related experiences such as Dependency and Recovery), and Objectives (detailing the aspects shared within posts related to SUD experiences). After testing and evaluating our new taxonomy, we annotated posts and developed multi-level and multi-label classification models using different baseline and state-of-the-art approaches. These models were used to categorize different types of personal drug experiences shared in posts from four prominent drug-related subreddits *r/cocaine*, *r/opiates*, *r/stims*, *r/benzodiazepines*. These subreddits were selected as they are the largest subreddits that represent substances identified as most commonly abused (NIDA, 2023). The results of our analysis show that GPT-4 outperformed other models and more accurately labeled classes across each of the three levels. Further exploration of the labeled posts showed that online posts are more inquisitive in nature and discuss themes relating to desired or undesired effects of the substance and/or how to consume them. Especially in the context of Recovery-related posts, users emphasize themes of Nurturant support, Relapse, and Safety. Moreover, our results show that discussions around Dependency frequently cover the Effects and Methods of Ingestion. Psycholinguistic analysis revealed that posts that contain topics such as Safety, Combination of Substances, and Effects tend to share personal experiences and exhibit a higher prevalence of language indicative of harm reduction efforts, as well as personal disclosures on family and social support systems.

Our study makes several contributions. Firstly, it introduces a new taxonomy for the classification of personal drug experiences. Additionally, we have developed a human-annotated dataset comprising 500 Reddit posts related to drug use, which showcases the wide range of personal drug experiences discussed in user-generated content. We also demonstrate the capability for automatic classification of personal drug experiences across three levels and multiple classes. Lastly, our work includes an analysis of the narratives surrounding substance use, SUD, and recovery based on self-disclosed user experiences. This work enriches the understanding of PWUD's experiences and informs the broader knowledge base on SUD and drug use.

## 2 Related Work

### 2.1 Exploring Personal SUD Narratives

Within the two drug-related subreddits, *r/trees* and *r/opiates*, Costello et al. (2017) utilized hermeneutic content analysis to categorize forum posts into eight distinct groups: disclosure, instruction, drug culture, community norms, moralizing, legality, and banter and identified three primary motives for PWUD to share information: to offer advice, seek information from others, and provide context for illicit disclosures. Wombacher et al. (2020) applied content analysis to the well-known drug subreddit, *r/Drugs*, to identify the types of social support exchanged among active substance users and found that the majority of the support was action-facilitating, aimed at safer drug use, with emotional support also being significant.

Recent studies used NLP and machine learning to analyze drug-related discussions on social media, offering insights into user behavior and content shared. Balani and De Choudhury (2015) analyzed over 30,000 posts from mental health-related subreddits, and found a significant amount of self-disclosure among users, with those engaging more intensely showing longer activity on the platform. Strapparava and Mihalcea (2017) used mental health forum posts to classify DSM-5 categories via zero-shot learning, employing n-grams and LDA topics. Incorporating slang improved accuracy by reducing false alarms by 17%, and domain knowledge enhanced recall. Varma et al. (2022) applied few-shot learning models to identify suicide risk on Reddit and found that few-shot learning with outlier removal and Support Vector Machine classifier yields better accuracy in detecting suicide risk.

Our study extends prior work by incorporating

insights, taxonomies, and methods from existing research on online communications. We developed a refined and extensive taxonomy for analyzing user-generated texts, alongside a codebook for human annotation. This new taxonomy aims to provide a theoretical foundation, ensuring accuracy and consistency in annotating user interactions.

### 2.2 LLM-based Information Extraction and Annotation

In recent years, the rapid advancements of LLMs such as GPT models, LLaMA, OPT, and BLOOM, have facilitated various tasks in NLP (Wei et al., 2022a), demonstrating great performance across a wide range of NLP tasks such as question answering (Trivedi et al., 2022), named-entity recognition (Wang et al., 2023), dialogue (Thoppilan et al., 2022), translation (Peng et al., 2023), and emotion analysis (Lei et al., 2023), often outperforming other models in zero-shot and few-shot contexts. LLMs' in-context learning (ICL) and text classification capabilities enable generating prompt-based textual responses, often with minimal examples (Wei et al., 2022b; Sun et al., 2023). Their proficiency in mimicking human text and labeling facilitates scalable NLP tasks in information retrieval with context sensitivity (Li et al., 2023a). Brown et al. (2020) demonstrated a few-shot classification method with GPT-3 for various NLP tasks, using text-based task definitions and demonstrations. The model showcased proficiency in on-the-fly reasoning and domain adaptation tasks like word unscrambling, novel word usage, and 3-digit arithmetic.

LLMs have been applied in human-in-the-loop and co-annotation methods (Li et al., 2023b). Cost efficiency is improved by using consensus methods between human and AI outputs (Chaganty et al., 2018; Zhang et al., 2021). Kang et al. (2023) explored combining LLM distillation with manual annotation, while Wang et al. (2021) looked into active labeling via logit outputs. While co-annotation balances quality and cost by leveraging uncertainty measures and evaluation thresholds (Li et al., 2023b), it faces challenges, as LLMs inherently struggle with generating structured abstract meaning, necessitating additional adaptation techniques (Ettinger et al., 2023).

The use of LLMs in the healthcare domain has shown promise in expanding the capacity of tasks such as summarization of patients' health records or question-answering (chatbot) (Liu et al., 2023; Nov et al., 2023). For instance, Garg et al. (2021) used LLMs to detect patient deterioration from electronic health records, demonstrating the potential of LLMs to accurately identify critical events. Similarly, Rijen et al. (2019) utilized LLMs to classify online health forum posts, demonstrating their ability to handle free-text data. Drawing on insights from previous studies, we employed a range of models, including LLM, to extract domain-specific information and classify personal drug experiences in user-generated content.

## 3 Data & Taxonomy Development

### 3.1 Data Collection

Data for this study was obtained using the Reddit API and the Python for Reddit API Wrapper (PRAW).[1] We collected posts from four popular drug subreddits (*r/opiate*s, *r/benzodiazepines*, *r/stims*, and *r/cocaine*) posted between 2017 and 2022, resulting in collecting 267,748 posts. These subreddits were selected because they are the largest within their respective classes of commonly abused substances (NIDA, 2023). We developed the taxonomy and training/test sets by randomly sampling around 1,600 posts. Initially, 100 posts were divided into four sets of 25 for the preliminary labeling and taxonomy development. We then annotated 500 posts for model training and testing. An additional 1,000 posts were then selected for labeling using the best-performing model.

### 3.2 Taxonomy of Lived Experiences Online

To understand the nuanced experiences of PWUD, we develop a domain-specific taxonomy derived from analyzing user-generated texts to distill narratives and offer insights into the personal and social aspects of drug use on social media. To develop this taxonomy and a codebook for annotation, we employed a hybrid deductive-inductive approach. First, using the existing literature in the field of drug use and social media, we established a framework for understanding key themes and dimensions related to personal drug experiences. This deductive phase involved developing a set of codes based on concepts from the literature, such as the type or objectives of posts (giving or receiving advice) (Balani and De Choudhury, 2015; Valdez and Patterson, 2022), indication of social support (Gauthier et al., 2022), and recovery- or withdrawal-related discourse (D'Agostino et al., 2017). This

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

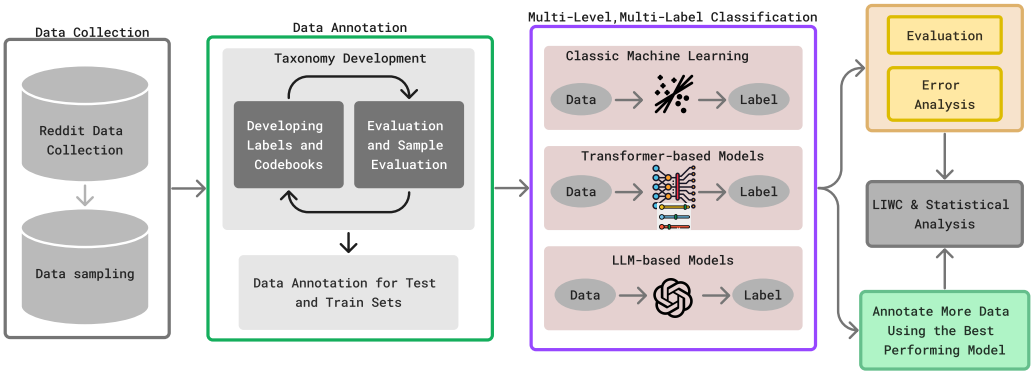



Figure 2: Classification and data analysis pipeline.

approach allowed us to create a comprehensive codebook tailored to our data and grounded in existing theories and research. Next, we applied these deductive codes to a sample of 25 Reddit posts. During coding, we asked our annotators (graduate and undergraduate students with a variety of experiences in health and computing) to identify new themes and concepts that were not covered by our initial codes. This resulted in the extraction of new concepts that were later added to the codebook as inductive codes. After coding a new sample of 25 posts, we reviewed and revised the codebook as necessary. This iterative process was repeated three times before we finalized the taxonomy and codebook, ensuring that it was well-suited to our data. In each iteration, annotators labeled a new set of posts in addition to those we labeled before to assess confidence in theme creation.

The final codebook consists of three levels: (1) Connection, (2) Subject, and (3) Objective. The type of Connection determines the overarching intent of the post, whether the user seeks to gain (Inquisition) or give information (Disclosure). Subject refers to the essence of the lived experience mentioned, particularly around Dependency, Recovery, and Other (typically recreational) experiences. While we labeled recovery and dependency-based discourse separately, we ensured that other circumstances were captured in our taxonomy. For example, posts referring to medicinal experiences (e.g., pain or anxiety) or an experience non-indicative of abuse were categorized as Other. Finally, the Objective captures the fine-grained topics discussed in the posts. For example, a post with a Recovery subject could include information on Safety, Quality, and Overdose. We consulted with domain experts in health to validate our new taxonomy, ensuring its efficiency and usefulness for professionals in the field. Table 4 in Appendix A provides an overview of the taxonomy for lived drug experiences, featuring definitions and examples for each code.

### 3.3 Annotation Process

After finalizing the codebook, three of the authors annotated two sets of 50 randomly selected posts. The agreement between each set of annotations was calculated ($k$ = .78 for Connection, and $k$ = .51 for Subject)[2], and any posts with disagreements were thoroughly discussed to establish a set of 100 mutually agreed-upon annotated posts. This dataset was used to test and evaluate the classification models. Once the annotators demonstrated a good understanding of the codebook, they annotated the 400 remaining posts. Regular check-ins were conducted between the annotators to ensure quality and consistency in using the codebook.

## 4 Multi-Level Lived Experience Detection

We used 400 annotated posts to train a series of multi-label models, each customized for different levels of classification granularity: Connection, Subject, and Objective. Minimal pre-processing was applied to maintain data integrity, which included converting text to lowercase, removing URLs, expanding contractions, and eliminating stopwords. We evaluate the models' performance on the test set consisting of 100 posts, using precision, recall, and F1 score metrics. Figure 2 shows the overall pipeline of this study.

**Baseline Models.** We implemented four baseline machine learning models to benchmark our drug experience classification system. We chose classic, feature-based algorithms for their interpretability

[2]Due to the complexity of the Objective level, despite multiple training sessions, there was still no substantial agreement between annotators across all 13 classes. With additional training, it may be possible to achieve higher inter-coder reliability.

and widespread use: Logistic Regression (LogR), Random Forest (RF), Support Vector Machines (SVM), and K-Nearest Neighbors (KNN). To provide a comprehensive evaluation, we trained and tested each model with two distinct feature sets: TF-IDF vectors and BERT (Devlin et al., 2019) pre-trained [CLS] token embedding. This approach allowed us to understand the impact of feature representation on model performance. In addition, we integrated OpenAI's `ada-002` architecture into our approach. Leveraging the capabilities of these LLM, we generated high-level features and trained our selected classic models.

**Transformer-based Models.** We used pre-trained language models from Hugging Face (Wolf et al., 2020) to extract rich contextual features. Specifically, BERT-base-uncased (Devlin et al., 2019), RoBERTa-base (Liu et al., 2019), DeBERTa-base (He et al., 2021), and BioBERT-v1.1 (Lee et al., 2020) were used. All models were fine-tuned and trained using a cross entropy loss function for 5 epochs and a learning rate of $2e-5$. During training, we optimized the model with AdamW optimizer(Loshchilov and Hutter, 2018).

Recognizing the limitations of large-scale annotation, we explored few-shot learning techniques. We employed SetFit (Tunstall et al., 2022), one of the most efficient few-shot architectures, leveraging SentBERT (Reimers and Gurevych, 2019) features. This choice aligns with current research advocating for few-shot models in situations where extensive annotation is impractical or costly (Schick and Schütze, 2021). To optimize performance, we fine-tuned the model using the cosine similarity loss function, for 2 epochs with a learning rate of $2e-5$ and a batch size of 8.

**LLM-based Models.** To assess the effectiveness of LLMs in classifying texts within the domain of SUD, we conducted a comprehensive experiment involving three models: two proprietary models from OpenAI, GPT-3.5 (gpt-3.5-turbo-0613) and GPT-4 (gpt-4-0125-preview) (OpenAI, 2023), and one open-source model, Mixtral 7B (Jiang et al., 2024). We evaluated their performance across three prompting styles:

- **Instruction-only ('I'):** This approach provided only the classification task instructions to the LLM.
- **Instruction & Definition ('I + D'):** Instructions with clear definitions of each classification label.

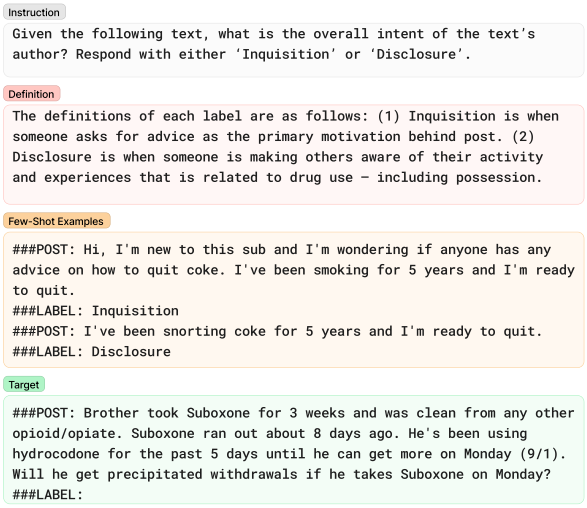



Figure 3: LLM prompting for Type of Connection

- **Instruction, Definition, & Examples ('I + D + E'):** Supplementing the definitions with two relevant examples for each label, mimicking few-shot learning scenarios.

Figure 3 illustrates the style of prompting used to label each post in our dataset. The dataset and models are available in our GitHub repository.[3]

## 5 Experiments Results

### 5.1 Data Analysis

We present the distribution of classes in Table 1, which was derived by applying our taxonomy to 500 posts. Between the testing and training sets, approximately 66.2% (N = 331) of the type of Connection were labeled as Inquisition, and 61% were labeled as Other for Subject. Among the Objectives, Methods of Ingestion (N = 265, 53%) and Effects (N = 251, 50.1%) were the most prevalent, while Overdose (N = 9, 1.8%) was the least prevalent.

### 5.2 Classification

Table 2 summarizes a selected set of results of our experiments using (1) baseline models, (2) Transformer-based models, and (3) LLMs-based models. As shown in the table, for the binary task of Connection (Inclusion vs. Disclosure), DeBERTa outperformed all models with respect to precision at 0.95, indicating its effectiveness in returning more relevant results overall. Using a Few-Shot model did not increase the performance compared to the baseline. However, the GPT-4

[3] https://github.com/social-nlp-lab/Drug-experience-classification

| | | Connection | | Subject | | | Objectives | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | WC | Inquisition | Disclosure | Dependency | Recovery | Other | C.o.S | E | L | M.H | M.o.I | N.S&M | O | Q | R | S | W |
| Testing | 106.1 | 51 | 49 | 37 | 7 | 53 | 20 | 66 | 5 | 24 | 66 | 11 | 5 | 9 | 7 | 26 | 13 |
| Training | 314.5 | 280 | 118 | 116 | 7 | 256 | 77 | 185 | 12 | 38 | 199 | 37 | 4 | 32 | 14 | 55 | 37 |
| Total | 147.7 | 331 | 167 | 153 | 14 | 309 | 97 | 251 | 17 | 62 | 265 | 48 | 9 | 41 | 21 | 81 | 50 |

Table 1: Frequencies of classes in each level labeled within training and test sets. Abbreviations: WC: word count; C.o.S: combination of substances; E: effects; L: legality; M.H.: mental health; M.o.I: methods of ingestion; N.S&M: nurturant support & morality; O: overdose; Q: quality; R: relapse; S: safety; W: withdrawal.

| | | | Connection | | | Subject | | | Objective | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Feature | Classifiers | Prec | Rec | F1 | Prec | Rec | F1 | Prec | Rec | F1 |
| Baseline | TF-IDF | KNN | 0.77 | 0.72 | 0.68 | 0.49 | 0.51 | 0.49 | 0.35 | 0.44 | 0.39 |
| | BERT | LogR | 0.71 | 0.72 | 0.71 | 0.53 | 0.58 | 0.55 | 0.60 | 0.47 | 0.49 |
| | BERT | SVM | 0.75 | 0.75 | 0.75 | 0.56 | 0.60 | 0.58 | 0.52 | 0.45 | 0.44 |
| | ada-002 | LogR | 0.82 | 0.82 | 0.81 | 0.57 | 0.63 | 0.58 | 0.48 | 0.30 | 0.34 |
| Transformer-Based | DeBERTa | DeBERTa | **0.95** | 0.87 | 0.90 | 0.66 | 0.71 | 0.69 | 0.51 | 0.44 | 0.43 |
| | RoBERTa | RoBERTa | 0.85 | 0.81 | 0.83 | 0.63 | 0.69 | 0.66 | 0.39 | 0.43 | 0.41 |
| | - | SetFit | 0.90 | 0.64 | 0.75 | 0.62 | 0.39 | 0.45 | 0.37 | 0.43 | 0.39 |
| LLM | - | GPT4 I + D | 0.76 | 0.45 | 0.32 | **0.77** | 0.64 | 0.66 | 0.56 | **0.76** | 0.61 |
| | - | GPT4 I + D + E | 0.91 | **0.91** | **0.91** | 0.74 | **0.72** | **0.73** | **0.71** | 0.66 | **0.64** |
| | - | Mixtral I + D + E | 0.78 | 0.51 | 0.43 | 0.72 | 0.34 | 0.23 | 0.42 | 0.84 | 0.53 |

Table 2: Weighted Precision, Recall, F1 scores for each level of experience for performant classification models. The best score for each metric is in **bold**. Metrics for all experiments can be found in Appendix 5.

model with 'I + D + E' achieved the highest Recall (0.91) and F1 scores (0.91) among all models.

We observed a similar trend in the Subject level, which includes more fine-grained labels (i.e., Dependency, Recovery, Other, and N/A). Specifically, the GPT-4 model using the 'I + D + E' method achieved the highest recall (0.72) and F1 score (0.73), while the GPT-4 model with the 'I + D' strategy outperformed the other models in terms of precision (0.77), with around 5% improvement.

Similarly, for the Objective classes, the best performing model was GPT4 with 'I + D + E' with respect to precision (0.71) and F1 (0.64). However, the GPT4 model with 'I + D' outperformed other models in recall (0.76), suggesting its ability to correctly identify a higher proportion of relevant instances. A more detailed discussion of the classification procedure and results is provided in Appendix B and Table 5.

## 6 Usefulness of our Models

In this study, we introduce a new taxonomy to better understand the lived experiences of PWUD online and analyze these experiences in user-generated texts on Reddit. Our study builds upon existing literature that employed qualitative analysis to uncover elements of social support within drug-focused Reddit communities (Bunting et al., 2021; Gauthier et al., 2022; D'Agostino et al., 2017; Graves et al., 2022; Wombacher et al., 2020). However, we apply computational methods to conduct this exploration, a novel approach compared to previous studies that have primarily concentrated on binary classifications (Al-Garadi et al., 2021) or other aspects of mental health (Balani and De Choudhury, 2015; Gaur et al., 2018; Valdez and Patterson, 2022). Our experiments demonstrate significant performance differences between classical machine learning, state-of-the-art transformer-based models, and LLMs for multi-label classification of drug-related posts. LLMs outperformed other models in detecting diverse aspects of lived experience disclosures, underscoring their potential for larger-scale investigations into personal narratives of drug use.

Current U.S. addiction recovery frameworks often fail to consider the lived experiences of individuals with SUD (Lipari et al., 2016), assuming universal access to professional guidance—a reality not shared by many affected. Online communities fill this gap, offering acceptance, understanding, and validation not always present in professional settings, thus becoming crucial for those lacking healthcare access (Mead and MacNeil, 2006). They enable discussions on substance use and recovery,

playing a vital role for individuals isolated from traditional healthcare resources(Boisvert et al., 2008). Our study aims to address these knowledge gaps by detailing the experiences of PWUD, enhancing the ability of experts to offer more comprehensive and personalized support. Our results show a balanced mix of users seeking (Inquisition) and sharing (Disclosure) information, aligning with prior work (Valdez and Patterson, 2022). Analyzing self-disclosed experiences within the posts indicates a minor portion discussing recovery experiences, touching on Nurturant Support, Relapse, and Safety. This finding is expected as recovery subreddits like *r/OpiatesRecovery*, and *r/benzorecovery*, host more detailed recovery discussions.

Prior research identified nurturant support themes of recovery and addiction in the *r/Drugs* subreddit(Wombacher et al., 2020). Our study builds on this, offering deeper insights into the nuanced discussions of recovery and dependency, as shown in this post:

*"171 days sober, I last posted 140 days ago about how getting cleaned up from opiates has revolutionized my life. just wanted to check in with everyone and see how everyone's doing."*

Among posts that disclosed a Dependency experience (N = 153), most contained themes of Effects and Methods of Ingestion, resonating with the concept of action-facilitating support identified in the *r/Drugs* community (Wombacher et al., 2020):

*"... I was on opiates. I started at ten-mg hydro twice a day for 2-3 months, after being directed to a pain doctor they were able to move me to 10mg/325 Percocet 4x a day since July...With all the horrible stigma surrounding opiates, I'm anxious to talk to my doctor about bumping or changing my meds...My tolerance has gone...I sometimes notice WD symptoms when I wait too long"*

This relationship confirms the value of online forums as spaces for sharing and obtaining information on substance use and recovery, highlighting our study's relevance. Moreover, examining conversations about Ingestion Methods and Effects can aid harm reduction. Sharing insights on safer practices, dosages, and possible side effects can help users reduce substance use risks.

### 6.1 Error Analysis

To gain a more comprehensive understanding of the best model's performance, we conducted a detailed error analysis. Despite the notable performance of GPT-4 'I + D + E' in classifying a broad range of classes across all three levels, we found some misclassified instances in the test set. For posts tagged as Connection, out of 22 tagged as Disclosure, only 3 were incorrectly classified as Inquisition, and from the 31 posts tagged as Inquisition, 2 were misclassified as Disclosure. At the Subject level, of the 17 posts tagged as Dependency, 4 were incorrectly classified as Recovery. Examples of these misclassified instances are presented in Table 8.

Our analysis shows that the model sometimes had difficulty discerning the subtle intentions expressed in personal narratives about drug experiences. This was particularly true in instances where narratives combined Disclosure (sharing of personal substance use experiences) with a question for feedback or solutions, not directly related to SUD. The dual nature of these communications often made it challenging for the model to determine the dominant intent behind the posts. Further analysis at the Subject level identified cases where the model incorrectly classified posts revolved around Dependency as Recovery. In some instances, even brief mentions of sobriety within an individual's narrative were interpreted by the model as signs of Recovery. This underscores the model's tendency to occasionally misjudge the context or importance of specific keywords in the discourse. Furthermore, our findings indicate a discrepancy in the Objective level classification, stemming from the model's inclination to over-rely on specific keywords while neglecting the wider context, potentially skewing the narratives. This skewed portrayal of user narratives underscores the need for domain experts to be involved in promoting and evaluating models, thereby enhancing the model's ability to understand and interpret context more effectively.

## 7 The Language of Personal Drug Experiences

We applied our most effective model, GPT4 'I + D + E', to extend our analysis to 1,000 randomly selected posts from our dataset for psycholinguistic analysis of PWUD's lived experiences. To ensure a more representative sample and increase statistical power, we augmented the pre-annotated set of 500 posts with 1,000 randomly selected posts. This resulted in a comprehensive analysis of 1,500 posts across three dimensions: Connection, Subject, and Objective. Table 6 presents the distribution of labels across our dataset. We applied the Linguistic Inquiry and Word Count (LIWC) tool

| | $\mu_{Inquisition}/\mu_{Disclosure}$ | | | | | | | | | | | $\mu_{Recovery}/\mu_{Dependency}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| LWC | Recovery | Dependency | Effects | Methods of Ingestion | Comb. of Substances | Mental Health | Nurturant & Morality | Withdrawal | Safety | Relapse | ALL | Effects |
| **Qmark** | 6.39 | 5.19 | 1.15 | 1.28 | 0.85 | 0.94 | 2.79 | 5.03 | 0.86 | 3.48 | 2.70 | |
| **illness** | | | 3.03 | | 3.03 | | | | 2.97 | | | |
| **feeling** | | | | | 2.56 | | | | | | | |
| **risk** | | | | | 1.98 | | | | 1.98 | | | |
| **tentat** | | 1.73 | 1.97 | 2.14 | 2.24 | 1.44 | | | 2.21 | | 2.08 | |
| **curiosity** | | | | 1.33 | | | | | 1.39 | | 2.02 | |
| **discrep** | | 1.39 | 1.27 | 1.62 | 1.87 | 1.74 | | | 1.52 | | 1.59 | |
| **polite** | | 2.39 | 2.23 | 2.70 | 7.61 | 5.80 | | | 4.75 | | 1.51 | 1.72 |
| **cause** | | 1.82 | 1.40 | 1.54 | 1.73 | 1.20 | | | 1.47 | | 1.51 | |
| **insight** | | 1.42 | 1.12 | 1.21 | 0.93 | 1.25 | | | 0.99 | | 1.48 | |
| **emo_sad** | | | | | | | | | | | 0.86 | 3.43 |
| **tone_pos** | | | | | | | | | | | 0.58 | |
| **money** | | | | | | | | | | | 0.57 | |
| **home** | | | 0.23 | 0.25 | 0.09 | | | 0.12 | | | 0.50 | 2.92 |
| **we** | | | 0.37 | | | | | 0.09 | 0.09 | | 0.49 | |
| **family** | | 0.72 | 0.60 | 0.60 | | | | 0.57 | | | 0.47 | 1.49 |
| **affiliation** | | | | | | | | 0.45 | | | | |
| **moral** | | | | | | | | | | | 0.42 | |
| **emo_pos** | | | 0.38 | | | | | | | | 0.41 | |
| **memory** | | | 0.38 | | | | | | | | 0.40 | |
| **reward** | | | | | | | | | | | 0.39 | 5.71 |
| **wellness** | | | | | | | | | | | 0.30 | 3.74 |

Table 3: Top 20 largest and smallest LIWC categories by effect size. Highlighted cells indicate different ratios: dark blue represents the lowest ratios (smaller effect sizes), and dark red represents the highest ratios (larger effect sizes). Empty cells denote non-significant results.

(Boyd et al., 2022), which quantifies the prevalence of words from diverse categories in the text, to this dataset. To compare the average scores for LIWC categories, we conducted non-parametric Mann-Whitney U tests (Mann and Whitney, 1947), identifying statistically significant differences (with $p < .05$) between Inquisition vs. Disclosure and Recovery vs. Dependency posts. The Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) was employed to control the false discovery rate across 85 selected LIWC categories. Our analysis of posts considered two metrics: (1) the ratio for posts labeled as Recovery and Dependency ($R_R = \mu_{Recovery}/\mu_{Dependency}$), and (2) the ratio of mean LIWC category scores between Inquisition and Disclosure posts ( $R_I = \mu_{Inquisition}/\mu_{Disclosure}$). Detailed results can be found in Appendix D. We found that 47 LIWC categories were significantly associated with Inquisition posts ($p < .05$) overall, while six were significantly associated with Recovery posts. Table 3 presents the LIWC categories with the largest and smallest effect sizes (i.e., the ratios farthest from 1).

**Inquisition vs. Disclosure.** As expected, posts seeking information (Inquisition) used significantly more question marks (*Qmark*). However, this pattern was less pronounced in posts discussing Safety (0.94), Combination of Substances (0.84), and Mental Health (0.86), suggesting users disclose more information about these topics. Notably, we observed 1.16 times greater *authenticity* in Inquisition posts compared to Disclosure posts. Posts discussing Withdrawal experiences were less inquisitive, indicating a preference for sharing personal experiences rather than seeking information on this sensitive topic. Higher usage of LIWC categories such as *friend*, *conflict*, *affiliation* further support this interpretation (Table 7).

The use of *prosocial* language (e.g., "care", "help") was 2.54 times more frequent in posts inquiring about Dependency concerns. This may suggest that individuals seeking information about dependency may frame their inquiries in a way that elicits support and empathy. We found *prosocial* language also more prevalent in Disclosure posts related to Combination of Substances (0.46), indicating that discussions of polysubstance use may involve a greater degree of mutual support. While the use of language to establish *Clout* was generally more prevalent in Disclosure posts (0.89), likely reflecting a desire to establish credibility when sharing personal experiences, this pattern was particularly notable in discussions of Methods of Ingestion (0.69), suggesting that individuals sharing experiences related to drug use methods may be especially motivated to present themselves

as knowledgeable and authoritative.

Finally, while our analysis revealed significant differences in the use of *relig* and *food* categories, further investigation highlighted the limitations of LIWC limitations in capturing the nuances of drug-related discourse. These terms often appear in metaphorical contexts within drug discourse (e.g., "god," "hell," "cook," "bake"), underscoring the need for specialized linguistic tools tailored to the unique language of online drug forums.

**Combination of Substances and Safety.** Posts inquiring about combining substances used significantly more *polite* language (7.6 times more) than those disclosing experiences. This suggests a strategic use of politeness to foster a supportive environment when seeking potentially sensitive information. More frequent use of *acquire* language (e.g., "get," "take") in these inquiries could reflect a desire for information and a potential interest in consuming multiple substances. The threefold increase in *illness* language within posts about physiological Effects, Combination of Substances, and Safety suggests that users are prioritizing harm reduction and seeking information about potential negative health consequences. The significant association of *risk* language specifically with posts about combining substances and Safety underscores the growing concern users may have regarding the potential dangers of polysubstance use.

**Discussion of Physiological Effects.** The use of *polite* language is approximately 30% more frequent in Inquisition posts about physiological Effects (ratio of 2.23) compared to Recovery posts (ratio of 1.72). This difference may reflect a greater degree of deference or caution when seeking information about potentially stigmatized experiences. On the other hand, the more prevalent use of *emo_sad* language in Recovery posts about effects (3.4 times more than in Dependency posts) suggests that individuals sharing personal experiences related to recovery may be more likely to express negative emotions associated with the physiological consequences of drug use. The increased use of *home* (e.g., "home," "bed") and *family* (e.g., "mother," "brother") language in Recovery posts, which were 2.9 and 1.49 times more frequent respectively, suggests that discussions of recovery often involve reflections on personal relationships and living environments, potentially highlighting the importance of social support and stable environments in the recovery process. The higher use of *reward* and *wellness* language in Recovery posts compared to Dependency posts suggests a focus on positive outcomes and potential benefits of overcoming physiological dependencies.

## 8 Conclusion

This study aims to identify how individuals discuss their personal drug experiences online. Using a deductive-inductive approach, we developed a taxonomy to assess user-generated posts, including their intended Connections (Inquisition or Disclosure), Subjects (e.g., Recovery, Dependency, Other), and specific Objectives (e.g., Relapse, Quality, Safety, Legality). We then employed this taxonomy to annotate 500 randomly sampled posts from a dataset we created, consisting of posts from four subreddits: *r/opiates*, *r/benzodiazepines*, *r/stims*, and *r/cocaine*. We used this data to train three sets of classifiers: (1) baseline models, (2) transformer-based deep learning models, and (3) LLM-based models including GPT-3.5 Turbo, GPT-4, and Mixtral, an open-source LLM. Our data analysis shows that posts are more inquisitive and predominately revolve around Effects and Methods of Ingestion. Our classification results show that GPT-4 with Instruction, Definition, and Examples ('I +D+ E') in prompts outperformed other models, whereas the DeBERTa-based transformer model had the best performance among the non-LLM models.

After applying our best-performing model to an additional 1,000 randomly selected posts, we used LIWC to analyze the linguistic differences between posts labeled as Inquisition vs. Disclosure and Recovery vs. Dependency. This analysis revealed that Inquisition posts significantly used more *authenticity*-related language, while Disclosure posts emphasized personal sharing, especially in sensitive topics like Withdrawal. The analysis also highlighted the varied use of *prosocial*, *clout*, and metaphorical language across different discussion themes, providing deeper insights into the psychosocial dynamics of online drug-related discourse.

These findings provide insight into the intricate language of drug use discussions, highlighting potential indicators of SUD or recovery initiation. Our results underscore how online forums provide crucial support and safety planning, demonstrating the value of computational analysis in understanding health-related online communities and informing SUD treatment and recovery interventions.

## Acknowledgements

We would like to thank Dr. Robert Sterling, Lauren Kairys, and Maggie Dickinson for their assistance with evaluating our codebook and their insightful feedback. We thank Sanonda Gupta for her help and Satvik Bahsin, Amui Gayle, Donald Hattier, Lauren Miller, Linh Nguyen, and Medhavi Pandit for all their hard work in the development of the codebook used for annotation. We also want to acknowledge the Reddit users whose discussions on various aspects of drug use served as the backbone of this study.

## 9 Limitations

Thematic analysis, inherently subjective, depends on the annotator's interpretation, potentially introducing biases into the coding process and affecting data accuracy. Focusing solely on four subreddits to derive insights into the SUD and recovery communities brings limitations, given the existence of numerous forums on these topics. Our selection aimed to reflect the broader nuances within the SUD landscape, acknowledging the constraints of such a scope.

Human annotation, while detailed, poses challenges due to its time-intensive nature for large datasets, limiting the data volume for model training and possibly affecting the models' performance and applicability. Moreover, the efficacy of few-shot learning is contingent upon the quality and diversity of training data; limitations include the risk of overfitting, computational inefficiencies, and poor generalization to new tasks or data distributions in cases of noisy or insufficient data.

Annotator unfamiliarity with specific substances could lead to misinterpretations that experts might avoid, underscoring the models' goal to grasp context without always requiring domain expertise. Additionally, the predominantly pseudonymous nature of Reddit participation, skewing towards a younger, majority male demographic, coupled with Hargittai (2020) observation that social media users tend to be of higher socioeconomic status, suggests a potential skew in the perspectives represented in computational social science research.

Our classification models excel in different levels of categorization but struggle with fine-grained, domain-specific classes, leading to potential misclassifications or oversimplifications. This limitation poses a challenge to the accurate recognition of nuanced distinctions within posts. Enhancing the models' capabilities in identifying these detailed classes (e.g., by using external knowledge bases) is essential for improving the utility of our models in domain-specific applications, thereby enabling more precise analysis and targeted interventions. Future work will focus on addressing these limitations.

## 10 Ethics Statement

In alignment with the harm reduction perspective, which prioritizes the autonomy and well-being of PWUD (Coalition), our research methodology places significant importance on protecting the privacy and anonymity of Reddit users who share their personal experiences. This study has received approval from the University's Institutional Review Board (IRB), ensuring adherence to rigorous ethical standards. To further safeguard user privacy, we will not release the full text of Reddit posts. Instead, we will only make available post IDs alongside their corresponding labels, allowing other researchers with appropriate access to reproduce our findings while ensuring user anonymity. Additionally, any quotes used in this work have been carefully modified to remove identifying details and protect user privacy. This involved paraphrasing content, removing specific drug names or slang terms, and generalizing language used in the quotes. We believe this approach upholds the principles of harm reduction while enabling valuable research to be conducted in a manner that respects the dignity and confidentiality of the individuals who generously share their experiences.

## References

Mohammed Ali Al-Garadi, Yuan-Chi Yang, Haitao Cai, Yucheng Ruan, Karen O'Connor, Gonzalez-Hernandez Graciela, Jeanmarie Perrone, and Abeed Sarker. 2021. Text classification models for the automatic detection of nonmedical prescription medication use from social media. *BMC Medical Informatics and Decision Making*, 21(1):27.

Sairam Balani and Munmun De Choudhury. 2015. Detecting and Characterizing Mental Health Related Self-Disclosure in Social Media. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '15, pages 1373–1378, New York, NY, USA. Association for Computing Machinery.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the*

*Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Rosemary A Boisvert, Linda M Martin, Maria Grosek, and Anna June Clarie. 2008. Effectiveness of a peer-support community in addiction recovery: participation as intervention. *Occupational Therapy International*, 15(4):205–220.

Layla Bouzoubaa and Rezvaneh Rezapour. 2024. Euphoria's hidden voices: Examining emotional resonance and shared substance use experience of viewers on reddit. In *Proceedings of the Workshop on Data for the Wellbeing of Most Vulnerable at the 18th International AAAI Conference on Web and Social Media (ICWSM)*, page 22. Association for the Advancement of Artificial Intelligence.

Layla Bouzoubaa, Jordyn Young, and Rezvaneh Rezapour. 2023. Exploring the landscape of drug communities on reddit: A network study. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, pages 558–565.

Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, pages 1–47.

Graham Brown, Sione Crawford, Gari-Emma Perry, Jude Byrne, James Dunne, Daniel Reeders, Angela Corry, Jane Dicka, Hunter Morgan, and Sam Jones. 2019. Achieving meaningful participation of people who use drugs and their peer organizations in a strategic research partnership. *Harm reduction journal*, 16(1):1–10.

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

Amanda M. Bunting, David Frank, Joshua Arshonsky, Marie A. Bragg, Samuel R. Friedman, and Noa Krawczyk. 2021. Socially-supportive norms and mutual aid of people who use opioids: An analysis of Reddit during the initial COVID-19 pandemic. *Drug and Alcohol Dependence*, 222:108672.

Arun Tejasvi Chaganty, Stephen Mussman, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evaluation. *arXiv preprint arXiv:1807.02202*.

Munmun De Choudhury and Sushovan De. 2014. Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):71–80. Number: 1.

National Harm Reduction Coalition. Principles of harm reduction.

Kaitlin L. Costello, III John D. Martin, and Ashlee Edwards Brinegar. 2017. Online disclosure of illicit information: Information behaviors in two drug forums. *Journal of the Association for Information Science and Technology*, 68(10):2439–2448.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805 [cs]*. ArXiv: 1810.04805.

Alexandra R. D'Agostino, Allison R. Optican, Shaina J. Sowles, Melissa J. Krauss, Kiriam Escobar Lee, and Patricia A. Cavazos-Rehg. 2017. Social networking online to recover from opioid use disorder: A study of community interactions. *Drug and Alcohol Dependence*, 181:5–10.

Allyson Ettinger, Jena D. Hwang, Valentina Pyatkin, Chandra Bhagavatula, and Yejin Choi. 2023. "You Are An Expert Linguistic Annotator": Limits of LLMs as Analyzers of Abstract Meaning Representation. ArXiv:2310.17793 [cs].

S. Garg, J. Taylor, M. El Sherief, E. Kasson, T. Aledavood, R. Riordan, N. Kaiser, P. Cavazos-Rehg, and M. De Choudhury. 2021. Detecting risk level in individuals misusing fentanyl utilizing posts from an online community on Reddit. *Internet Interventions*, 26.

Manas Gaur, Ugur Kursuncu, Amanuel Alambo, Amit Sheth, Raminta Daniulaityte, Krishnaprasad Thirunarayan, and Jyotishman Pathak. 2018. "Let Me Tell You About Your Mental Health!": Contextualized Classification of Reddit Posts to DSM-5 for Web-based Intervention. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 753–762, Torino Italy. ACM.

Robert P Gauthier, Mary Jean Costello, and James R Wallace. 2022. "I Will Not Drink With You Today": A Topic-Guided Thematic Analysis of Addiction Recovery on Reddit. *CHI Conference on Human Factors in Computing Systems*, pages 1–17. Conference Name: CHI '22: CHI Conference on Human Factors in Computing Systems ISBN: 9781450391573 Place: New Orleans LA USA Publisher: ACM.

Rachel Lynn Graves, Jeanmarie Perrone, Mohammed Ali Al-Garadi, Yuan-Chi Yang, Jennifer S. Love, Karen O'Connor, Graciela Gonzalez-Hernandez, and Abeed Sarker. 2022. Thematic Analysis of Reddit Content About Buprenorphine-naloxone Using Manual Annotation and Natural Language Processing Techniques. *Journal of Addiction Medicine*.

Eszter Hargittai. 2020. Potential biases in big data: Omitted voices on social media. *Social Science Computer Review*, 38(1):10–24.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced

bert with disentangled attention. (arXiv:2006.03654). ArXiv:2006.03654 [cs].

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Junmo Kang, Wei Xu, and Alan Ritter. 2023. Distill or annotate? cost-efficient fine-tuning of compact models. *arXiv preprint arXiv:2305.01645*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240. ArXiv:1901.08746 [cs].

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, and Sirui Wang. 2023. Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework. *arXiv preprint arXiv:2309.11911*.

Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. 2023a. Evaluating ChatGPT's Information Extraction Capabilities: An Assessment of Performance, Explainability, Calibration, and Faithfulness. ArXiv:2304.11633 [cs].

Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy F. Chen, Zhengyuan Liu, and Diyi Yang. 2023b. CoAnnotating: Uncertainty-Guided Work Allocation between Human and Large Language Models for Data Annotation. ArXiv:2310.15638 [cs].

Rachel N. Lipari, Eunice Park-Lee, and Struther Van Horn. 2016. America's need for and receipt of substance use treatment in 2015. *Substance Abuse and Mental Health Services Administration (SAMHSA)*.

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

Zhengliang Liu, Xiaowei Yu, Lu Zhang, Zihao Wu, Chao Cao, Haixing Dai, Lin Zhao, Wei Liu, Dinggang Shen, Quanzheng Li, et al. 2023. Deid-gpt: Zero-shot medical text de-identification by gpt-4. *arXiv preprint arXiv:2303.11032*.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

H. B. Mann and D. R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60.

Shery Mead and Cheryl MacNeil. 2006. Peer support: What makes it unique. *International Journal of Psychosocial Rehabilitation*.

NIDA. 2023. Commonly used drugs charts.

Oded Nov, Nina Singh, and Devin M Mann. 2023. Putting chatgpt's medical advice to the (turing) test. *medRxiv*, pages 2023–01.

OpenAI. 2023. Gpt-4 technical report.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of chatgpt for machine translation. *arXiv preprint arXiv:2303.13780*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. (arXiv:1908.10084). ArXiv:1908.10084 [cs].

Paul Van Rijen, D. Teodoro, Nona Naderi, Luc Mottin, J. Knafou, Matt Jeffryes, and Patrick Ruch. 2019. A Data-Driven Approach for Measuring the Severity of the Signs of Depression using Reddit Posts.

Timo Schick and Hinrich Schütze. 2021. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Carlo Strapparava and Rada Mihalcea. 2017. A Computational Analysis of the Language of Drug Addiction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 136–142, Valencia, Spain. Association for Computational Linguistics.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. *arXiv preprint arXiv:2305.08377*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *arXiv preprint arXiv:2212.10509*.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts. (arXiv:2209.11055). ArXiv:2209.11055 [cs].

Danny Valdez and Megan S. Patterson. 2022. Computational analyses identify addiction help-seeking behaviors on the social networking website Reddit: Insights into online social interactions and addiction support communities. *PLOS Digital Health*, 1(11):e0000143. Publisher: Public Library of Science.

Sandeep Varma, Shivam Shivam, Biswarup Ray, and Ankita Banerjee. 2022. Few shot learning with fine-tuned language model for suicidal text detectio. Technical report. Type: article.

Nora D. Volkow, Joshua A. Gordon, and George F. Koob. 2021. Choosing appropriate language to reduce the stigma around mental illness and substance use disorders. *Neuropsychopharmacology*, 46(13):2230–2232. Number: 13 Publisher: Nature Publishing Group.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. Gpt-ner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? gpt-3 can help. *arXiv preprint arXiv:2108.13487*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Kevin Wombacher, Sarah E. Sheff, and Natalie Itrich. 2020. Social Support for Active Substance Users: A Content Analysis of r/Drugs. *Health Communication*, 35(6):756–765. Publisher: Routledge _eprint: https://doi.org/10.1080/10410236.2019.1587691.

Yuan-Chi Yang, Mohammed Ali Al-Garadi, Jennifer S. Love, Hannah L. F. Cooper, Jeanmarie Perrone, and Abeed Sarker. 2023. Can accurate demographic information about people who use prescription medications nonmedically be derived from Twitter? *Proceedings of the National Academy of Sciences*, 120(8):e2207391120. Publisher: Proceedings of the National Academy of Sciences.

Jieyu Zhang, Bohan Wang, Xiangchen Song, Yujing Wang, Yaming Yang, Jing Bai, and Alexander Ratner. 2021. Creating training sets via weak indirect supervision. *arXiv preprint arXiv:2110.03484*.

## A Taxonomy of Personal Drug Experiences

As discussed in §3, following a comprehensive review of the literature and a detailed analysis of the posts in our dataset, we developed a taxonomy of lived drug experiences within online discussions. This taxonomy is structured across three levels: connection, subject, and objective. Through multiple iterations and evaluations, we refined our categorization to capture the nuances of online drug-related narratives accurately. Table 4 presents the rationale, detailed definitions, and examples for each level, offering insights into the complexity and diversity of these experiences shared online.

## B Details of Classification Procedure and Results

Table 5 presents the comprehensive list of classifiers and the corresponding results we obtained for the three levels of classification performed. The table encompasses a detailed comparison across different metrics, providing insights into the performance of each classifier within the context of our study. The classifications were conducted across three levels. For the baseline models, for both Connection and Subject levels, LogR and SVM models employ a one-versus-all strategy, whereas the KNN and RF models directly support multi-class classification. At the Objective level, however, we adopt the 'MultiOutputClassifier' strategy to address the multi-label classification task.

For the Objective level in transformer-based models, we adopted a multi-label classification strategy to manage samples that simultaneously belong to multiple categories.

The number of parameters in The BERT model is 110 million, while the RoBERTa and DeBERTa models have 125 million and 276 million parameters, respectively. We used the T4 GPU provided by Google Colab for our transformer-based models. For SentBERT model, we used seven few-shot samples from each category at both the Connection and Subject levels, and only four examples per category at the Objective level.

## C Multi-level, Multi-class Frequencies

Table 6 presents the distribution frequencies for each objective, categorized by connection and subject types. These results illustrate how the various classes are represented across a dataset of 1,500 entries, which includes 1,000 posts annotated using our best-performing model, complemented by an additional 500 manually annotated data points. This comprehensive overview aids in understanding the prevalence and patterns of different objectives within the context of the study, highlighting the robustness and coverage of our annotation approach.

## D LIWC and Statistical Analysis

Table 7 shows the ratios of mean scores for posts labeled as Recovery and Dependency ($\mu_{Recovery}/\mu_{Dependency}$) and the ratios of mean scores for Inquisition and Disclosure ($\mu_{Inquisition}/\mu_{Dependency}$) across different LIWC categories. LIWC assesses the text by calculating the proportion of words belonging to different psychologically relevant categories.

| Dimension | Rationale | Code | Definition and Example |
|---|---|---|---|
| Connection | What is the primary purpose of the post? Is the post asking for something from the community or is it meant to share stories and lived experiences? | Disclosure | Making others aware of the activity related to drug use, both primary and secondary accounts – including possession. Example: "I've been taking X for anxiety. I finally went to a doctor who told me he wouldn't prescribe me any benzo." |
| | | Inquisition | Asking for questions and advice on SUD. Example: "What is your cure to prevent coke side effects or things you do before, during, and after your session?" |
| Subject | What is the overarching subject of the post? | Dependency | The medical term used to describe drug or alcohol use that continues even when significant problems related to their use have developed. Example: "I'm 23 in usa been on drugs since I was a teen. drugs r bad but I chose to use, if you knew me know why!" |
| | | Recovery | Describing the process of overcoming substance use disorder and regaining physical, emotional, and mental health. Example: "I am sober for six weeks and it is like my life has changed completely." |
| | | Other | Other types of drug discussion NOT related to a recovery experience or indicative of Recovery nor Dependency. Can include general use. Example: "do you know of a drug that feels like X but wouldn't test +?" |
| | | N/A | Unrelated discussions, NOT related to any form of substance use. Example: "have you listened to Zoo Band?" |
| Objective | What are main themes or topic(s) corresponding to subject of the post? These themes provide information into users' motivation behind their post; including reasons for taking substances, the desired effects users hope to achieve, the benefits of using the substance, asking for support and encouragement. | Combination of Substances | The use of two or more substances at the same time. Example: "I'm taking 10mg of meth and 10mg X a day from the clinic. I'm on prescription, 10mg dex a day but sometimes go over." |
| | | Effects | Physical or emotional effect: The desired or undesired effects of the substance on the user's body and mind. Example: "I use amps for anxiety but now I have a bad craving." |
| | | Legality | The legal status of the substance in the user's jurisdiction. Example: "Is benzo legal in the US? Will I pass the test?" |
| | | Mental Health | The impact of the substance on the user's mental state, including mood, cognition, and emotional regulation. Example: "I feel like I have no options. The only thing kept me from feeling bad and having panic attacks or episodes of depression for extended periods are xanas." |
| | | Methods of Ingestion | The way in which the substance is consumed, such as smoking, inhaling, injecting, or swallowing. This encompasses different routes of administration and dosage-related considerations that play a crucial role in determining the substance's effects users. Example: "I'm smoking 3g coke, it's the way I like to use." |
| | | Nurturant Support & Morality | User's thoughts and feelings about the substance and their own use of it, including feelings of guilt, shame, or self-judgment. Nurturant support includes emotional, network, and esteem. Example: "I'm a horrible person because I am an addict, and I can't quit." |
| | | Overdose | The consumption of more of a substance than the body can safely handle, resulting in serious health problems or death. Example: "I overdosed last week, it was the scariest event of my life." |
| | | Quality | The purity or potency of the substance, or the user's perception of it. Example: "Got some prams that I'msure are pure, same size/dimensions/and weight." |
| | | Relapse | The return to using a substance after a period of abstinence. Example: "I relapsed on opiates last week, but I'm ok and in treatment and I want to stay sober." |
| | | Safety | The perceived or actual risk associated with using the substance, includes risks of overdose, addiction, or other negative consequences. Example: "Can my ex purposefully switch my syringe with one of hers because she is pissed about a comment I made about not risking Hep C?" |
| | | Withdrawal | The physical and psychological symptoms that occur when a person stops using a substance that they have been addicted to. Example: "I'm going through withdrawal now, it's awful." |
| | | Other | Any other objective that is not covered by the above categories. Example: "There was a post on various components of xanax and the meaning, it broke names into their various components." |
| | | N/A | Not related to SUD objective. Example: "Want to chat, hit me up." |

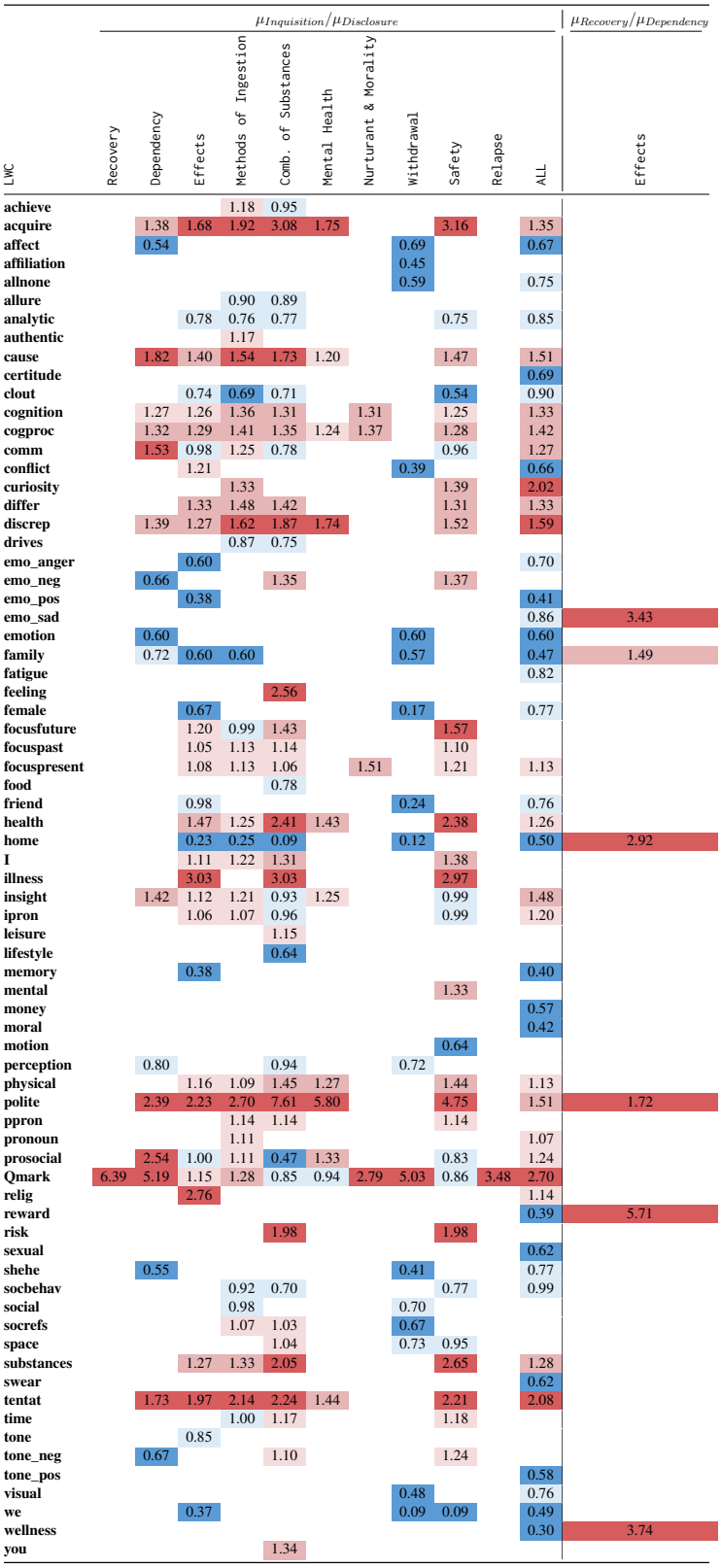
Table 4: Taxonomy of Lived Drug Experiences in Online Discussions

| | Feature | Classifiers | Connection | | | Subject | | | Objective | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Prec | Rec | F1 | Prec | Rec | F1 | Prec | Rec | F1 |
| Baseline | TF-IDF | KNN | 0.77 | 0.72 | 0.68 | 0.49 | 0.51 | 0.49 | 0.35 | 0.44 | 0.39 |
| | TF-IDF | SVM | 0.76 | 0.60 | 0.47 | 0.65 | 0.70 | 0.66 | 0.35 | 0.47 | 0.41 |
| | TF-IDF | RF | 0.67 | 0.66 | 0.62 | 0.52 | 0.58 | 0.55 | 0.37 | 0.32 | 0.35 |
| | TF-IDF | LogR | 0.76 | 0.74 | 0.72 | 0.58 | 0.53 | 0.52 | 0.46 | 0.47 | 0.41 |
| | BERT | LogR | 0.71 | 0.72 | 0.71 | 0.53 | 0.58 | 0.55 | 0.60 | 0.47 | 0.49 |
| | BERT | KNN | 0.67 | 0.62 | 0.53 | 0.55 | 0.53 | 0.52 | 0.36 | 0.45 | 0.40 |
| | BERT | RF | 0.39 | 0.51 | 0.42 | 0.55 | 0.60 | 0.58 | 0.45 | 0.39 | 0.40 |
| | BERT | SVM | 0.75 | 0.75 | 0.75 | 0.56 | 0.60 | 0.58 | 0.52 | 0.45 | 0.44 |
| | ada-002 | SVM | 0.8 | 0.8 | 0.8 | 0.26 | 0.53 | 0.35 | 0.33 | 0.29 | 0.30 |
| | ada-002 | LogR | 0.82 | 0.82 | 0.81 | 0.57 | 0.63 | 0.58 | 0.48 | 0.30 | 0.34 |
| | ada-002 | RF | 0.64 | 0.62 | 0.62 | 0.46 | 0.57 | 0.51 | 0.39 | 0.23 | 0.26 |
| | ada-002 | KNN | 0.78 | 0.78 | 0.78 | 0.46 | 0.57 | 0.51 | 0.39 | 0.23 | 0.26 |
| Transformer-Based | BERT | BERT | 0.78 | 0.68 | 0.73 | 0.58 | 0.56 | 0.54 | 0.38 | 0.41 | 0.41 |
| | DeBERTa | DeBERTa | **0.95** | 0.87 | 0.90 | 0.66 | 0.71 | 0.69 | 0.51 | 0.44 | 0.43 |
| | RoBERTa | RoBERTa | 0.85 | 0.81 | 0.83 | 0.63 | 0.69 | 0.66 | 0.39 | 0.43 | 0.41 |
| | BioBERT | BioBERT | 0.54 | 0.9 | 0.67 | 0.59 | 0.62 | 0.60 | 0.38 | 0.45 | 0.41 |
| | - | SetFit | 0.90 | 0.64 | 0.75 | 0.62 | 0.39 | 0.45 | 0.37 | 0.43 | 0.39 |
| LLM | - | GPT3.5 I | 0.6 | 0.6 | 0.6 | 0.73 | 0.45 | 0.4 | 0.57 | 0.61 | 0.56 |
| | - | GPT3.5 I + D | 0.49 | 0.43 | 0.39 | 0.66 | 0.47 | 0.42 | 0.54 | 0.57 | 0.53 |
| | - | GPT3.5 I + D + E | 0.71 | 0.7 | 0.69 | 0.75 | 0.68 | 0.68 | 0.61 | 0.69 | 0.62 |
| | - | GPT4 I | 0.73 | 0.57 | 0.53 | 0.74 | 0.34 | 0.23 | 0.54 | 0.74 | 0.6 |
| | - | GPT4 I + D | 0.76 | 0.45 | 0.32 | **0.77** | 0.64 | 0.66 | 0.56 | **0.76** | 0.61 |
| | - | GPT4 I + D + E | 0.91 | **0.91** | **0.91** | 0.74 | **0.72** | **0.73** | **0.71** | 0.66 | **0.64** |
| | - | Mixtral I | 0.34 | 0.58 | 0.43 | 0.1 | 0.32 | 0.16 | 0.4 | 0.68 | 0.48 |
| | - | Mixtral I + D | 0.17 | 0.42 | 0.24 | 0.59 | 0.43 | 0.37 | 0.54 | 0.68 | 0.55 |
| | - | Mixtral I + D + E | 0.78 | 0.51 | 0.43 | 0.72 | 0.34 | 0.23 | 0.42 | 0.84 | 0.53 |

Table 5: Weighted Precision, Recall, and F1 for all classification models.

| | | Comb. of Substances | Effects | Legality | Mental Health | Methods of Ingestion | Nurturant & Morality | Overdose | Quality | Relapse | Safety | Withdrawal | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Inquisition | Dependency | 27 | **88** | 3 | 34 | 61 | 19 | 11 | 12 | 14 | 32 | 26 | 328 |
| | Recovery | 2 | 9 | 1 | 7 | 5 | **17** | 1 | 1 | 15 | 3 | **17** | 78 |
| | Other | 31 | **119** | 18 | 28 | 87 | 18 | 8 | 20 | 4 | 40 | 3 | 397 |
| Disclosure | Dependency | 34 | 84 | 9 | 28 | **89** | 11 | 5 | 9 | 13 | 38 | 50 | 370 |
| | Recovery | 5 | 8 | 1 | 11 | 10 | 4 | 4 | 0 | 13 | 12 | **23** | 91 |
| | Other | 78 | 200 | 23 | 29 | **219** | 9 | 6 | 43 | 1 | 109 | 10 | 732 |
| Total | | 177 | 508 | 55 | 137 | 471 | 78 | 35 | 85 | 60 | 234 | 129 | |

Table 6: Frequencies of each objective with respect to each connection and lived experience types. The 'N/A' label for lived experience and 'N/A' and "Other' objectives labels are excluded from this table.



| LWC | $\mu_{Inquisition}/\mu_{Disclosure}$ | | | | | | | | | | $\mu_{Recovery}/\mu_{Dependency}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Recovery | Dependency | Effects | Methods of Ingestion | Comb. of Substances | Mental Health | Nurturant & Morality | Withdrawal | Safety | Relapse | ALL | Effects |
| **achieve** | | | | 1.18 | 0.95 | | | | | | | |
| **acquire** | | 1.38 | 1.68 | 1.92 | 3.08 | 1.75 | | | 3.16 | | 1.35 | |
| **affect** | | 0.54 | | | | | | 0.69 | | | 0.67 | |
| **affiliation** | | | | | | | | 0.45 | | | | |
| **allnone** | | | | | | | | 0.59 | | | 0.75 | |
| **allure** | | | | 0.90 | 0.89 | | | | | | | |
| **analytic** | | | 0.78 | 0.76 | 0.77 | | | | 0.75 | | 0.85 | |
| **authentic** | | | | 1.17 | | | | | | | | |
| **cause** | | 1.82 | 1.40 | 1.54 | 1.73 | 1.20 | | | 1.47 | | 1.51 | |
| **certitude** | | | | | | | | | | | 0.69 | |
| **clout** | | | 0.74 | 0.69 | 0.71 | | | | 0.54 | | 0.90 | |
| **cognition** | | 1.27 | 1.26 | 1.36 | 1.31 | | 1.31 | | 1.25 | | 1.33 | |
| **cogproc** | | 1.32 | 1.29 | 1.41 | 1.35 | 1.24 | 1.37 | | 1.28 | | 1.42 | |
| **comm** | | 1.53 | 0.98 | 1.25 | 0.78 | | | | 0.96 | | 1.27 | |
| **conflict** | | | 1.21 | | | | | 0.39 | | | 0.66 | |
| **curiosity** | | | | 1.33 | | | | | 1.39 | | 2.02 | |
| **differ** | | | 1.33 | 1.48 | 1.42 | | | | 1.31 | | 1.33 | |
| **discrep** | | 1.39 | 1.27 | 1.62 | 1.87 | 1.74 | | | 1.52 | | 1.59 | |
| **drives** | | | | 0.87 | 0.75 | | | | | | | |
| **emo_anger** | | | 0.60 | | | | | | | | 0.70 | |
| **emo_neg** | | 0.66 | | | 1.35 | | | | 1.37 | | | |
| **emo_pos** | | | 0.38 | | | | | | | | 0.41 | |
| **emo_sad** | | | | | | | | | | | 0.86 | 3.43 |
| **emotion** | | 0.60 | | | | | | 0.60 | | | 0.60 | |
| **family** | | 0.72 | 0.60 | 0.60 | | | | 0.57 | | | 0.47 | 1.49 |
| **fatigue** | | | | | | | | | | | 0.82 | |
| **feeling** | | | | | 2.56 | | | | | | | |
| **female** | | | 0.67 | | | | | 0.17 | | | 0.77 | |
| **focusfuture** | | | 1.20 | 0.99 | 1.43 | | | | 1.57 | | | |
| **focuspast** | | | 1.05 | 1.13 | 1.14 | | | | 1.10 | | | |
| **focuspresent** | | | 1.08 | 1.13 | 1.06 | | 1.51 | | 1.21 | | 1.13 | |
| **food** | | | | | 0.78 | | | | | | | |
| **friend** | | | 0.98 | | | | | 0.24 | | | 0.76 | |
| **health** | | | 1.47 | 1.25 | 2.41 | 1.43 | | | 2.38 | | 1.26 | |
| **home** | | | 0.23 | 0.25 | 0.09 | | | 0.12 | | | 0.50 | 2.92 |
| **I** | | | 1.11 | 1.22 | 1.31 | | | | 1.38 | | | |
| **illness** | | | 3.03 | | 3.03 | | | | 2.97 | | | |
| **insight** | | 1.42 | 1.12 | 1.21 | 0.93 | 1.25 | | | 0.99 | | 1.48 | |
| **ipron** | | | 1.06 | 1.07 | 0.96 | | | | 0.99 | | 1.20 | |
| **leisure** | | | | | 1.15 | | | | | | | |
| **lifestyle** | | | | | 0.64 | | | | | | | |
| **memory** | | | 0.38 | | | | | | | | 0.40 | |
| **mental** | | | | | | | | | 1.33 | | | |
| **money** | | | | | | | | | | | 0.57 | |
| **moral** | | | | | | | | | | | 0.42 | |
| **motion** | | | | | | | | | 0.64 | | | |
| **perception** | | 0.80 | | | 0.94 | | | 0.72 | | | | |
| **physical** | | | 1.16 | 1.09 | 1.45 | 1.27 | | | 1.44 | | 1.13 | |
| **polite** | | 2.39 | 2.23 | 2.70 | 7.61 | 5.80 | | | 4.75 | | 1.51 | 1.72 |
| **ppron** | | | | 1.14 | 1.14 | | | | 1.14 | | | |
| **pronoun** | | | | 1.11 | | | | | | | 1.07 | |
| **prosocial** | | 2.54 | 1.00 | 1.11 | 0.47 | 1.33 | | | 0.83 | | 1.24 | |
| **Qmark** | 6.39 | 5.19 | 1.15 | 1.28 | 0.85 | 0.94 | 2.79 | 5.03 | 0.86 | 3.48 | 2.70 | |
| **relig** | | | 2.76 | | | | | | | | 1.14 | |
| **reward** | | | | | | | | | | | 0.39 | 5.71 |
| **risk** | | | | | 1.98 | | | | 1.98 | | | |
| **sexual** | | | | | | | | | | | 0.62 | |
| **shehe** | | 0.55 | | | | | | 0.41 | | | 0.77 | |
| **socbehav** | | | | 0.92 | 0.70 | | | | 0.77 | | 0.99 | |
| **social** | | | | 0.98 | | | | 0.70 | | | | |
| **socrefs** | | | | 1.07 | 1.03 | | | 0.67 | | | | |
| **space** | | | | | 1.04 | | | 0.73 | 0.95 | | | |
| **substances** | | | 1.27 | 1.33 | 2.05 | | | | 2.65 | | 1.28 | |
| **swear** | | | | | | | | | | | 0.62 | |
| **tentat** | | 1.73 | 1.97 | 2.14 | 2.24 | 1.44 | | | 2.21 | | 2.08 | |
| **time** | | | | 1.00 | 1.17 | | | | 1.18 | | | |
| **tone** | | | 0.85 | | | | | | | | | |
| **tone_neg** | | 0.67 | | | 1.10 | | | | 1.24 | | | |
| **tone_pos** | | | | | | | | | | | 0.58 | |
| **visual** | | | | | | | | 0.48 | | | 0.76 | |
| **we** | | | 0.37 | | | | | 0.09 | 0.09 | | 0.49 | |
| **wellness** | | | | | | | | | | | 0.30 | 3.74 |
| **you** | | | | | 1.34 | | | | | | | |

Table 7: Summary of statistically significant LIWC categories related to **inquisition** using Mann-Whitney U test.$R = \mu_{Inquisition}/\mu_{Disclosure}$ between the means of inquisition and disclosure-related posts, and $R = \mu_{Recovery}/\mu_{Dependency}$ between the means of recovery and dependency-related posts. Highlighted cells represent different $R$ with dark blue representing the lowest and dark red representing the highest ratios. Empty cells denote non-significant results.

| Post | True Label | Predicted Label |
|---|---|---|
| I was convinced the police were monitoring my computer and were about to arrest me. I bought it from a very reputable person—I'll spare the details—but he's been well-known with a good reputation for years. I'm still feeling a bit disoriented after that, and I'm going to die going to work tomorrow can someone please let me know what the fuck went wrong there cheers | Disclosure | Inquisition |
| I quit for about a week due to a scheduled blood test and surprisingly, I didn't feel as awful as expected. I don't clearly remember that week, so I'm not certain how I felt. Later, I had my wisdom teeth removed and stopped again for a week before the procedure. During that time, the person I had been purchasing from unfortunately passed away. This was a big call and I have now been sober for about 3 months, with all that information how bad was I again, 46 120 lb male | Dependency | Recovery |
| I want to care, but reality hits and I just want it to disappear. Music and weed, are the answer to a question I can't even ask. I know I can find satisfaction in that combination I got to quit, bide me some time to figure out how to do it is that what I need to do | Dependency | Recovery |
| Now I'm just inquiring because of my paranoia about the prevalence of counterfeit medications these days. Are people really faking Etizest1 blister packs and boxes with no active ingredients or different, underdosed chemicals? I doubt they'd invest the effort to do that, considering some pharmaceutical companies in India manufacture our medicine, so there's definitely some level of quality control. | Safety | Legality, Quality, Safety, Mental Health |
| So, my usual dealer was out of supply for a few weeks, so they introduced me and my group of meth-using friends to another dealer. Turns out he had fire product and he is also an awesome person. | Quality | Quality, Nurturant support & Morality |

Table 8: Comparison of True and Predicted labels in some misclassified posts.