Words Matter: Reducing Stigma in Online Conversations about Substance Use with Large Language Models

Anonymous ACL submission

Abstract

Stigma is a barrier to treatment for individuals struggling with substance use disorders (SUD), which leads to significantly lower treatment engagement rates. With only 7% of those affected receiving any form of help, societal stigma not only discourages individuals with SUD from seeking help but isolates them, hindering their recovery journey and perpetuating a cycle of shame and self-doubt. This study investigates how stigma manifests on social media, particularly Reddit, where anonymity can exacerbate discriminatory behaviors. We analyzed over 1.2 million posts, identifying 3,207 that exhibited stigmatizing language towards people who use substances (PWUS). Using Informed and Stylized LLMs, we develop a model for de-stigmatization of these expressions into empathetic language, resulting in 1,649 reformed phrase pairs. Our paper contributes to the field by proposing a computational framework for analyzing stigma and destigmatizing online content, and delving into the linguistic features that propagate stigma towards PWUS. Our work not only enhances understanding of stigma's manifestations online but also provides practical tools for fostering a more supportive digital environment for those affected by SUD. Code and data will be made publicly available upon acceptance.

1 Introduction

Every day, people struggling with substance use disorders (SUD) face a pervasive and often hidden enemy: stigma. This stigma, often deeply ingrained in societal attitudes, can act as a significant barrier to treatment and recovery. In fact, only approximately 7% of people living with an SUD receive any form of treatment (Substance Abuse and Mental Health Services Administration, 2023), with stigma reported as a major barrier (Centers for Disease Control and Prevention, 2023). SUD is a critical public health challenge in the US and worldwide, and the substantial stigma associated with these conditions only exacerbates the problem. Traditional support systems, although beneficial, often remain underutilized due to their perceived inaccessibility or the overwhelming stigma surrounding SUD, thus rendering this topic a societal taboo.

Type	Statement
Original	I have no empathy for drug addicts. I had friends and family who have struggled with the "disease". Everyone knows what happens when you start, and you usually end up dead. Many of my old friends have become addicts and I don't understand especially the ones with kids.
De-stigmatized	I find it difficult to empathize with individuals facing substance use challenges. I had friends and family who encountered these difficulties. It's widely acknowledged that there are risks involved from the outset, and the outcomes are often heartbreaking. Several of my old friends have dealt with these challenges, and it's particularly perplexing to me when they are parents.

Table 1: Example of directed stigmatizing language. De-stigmatized version generated with our Informed + Stylized model using GPT-4 removed stereotypes and harmful context while preserving the tone (stigma is in red, destigmatized counterparts is in blue).

Social media platforms like Reddit have emerged as important spaces for community discussions (Bouzoubaa et al., 2023). However, the anonymity provided by these environments sometimes exacerbates stigmas, leading to discrimination. People suffering from SUD often encounter derogatory comments, judgment, or misinformation online (Schomerus et al., 2011), which can reinforce self-stigma and stop them from seeking help. The spread of stigmatizing attitudes on social media can also influence public opinion, further perpetuating the stereotypes and prejudices against those with SUD (McLaren et al., 2023). As a result, despite the potential for support, the digital space can mirror and magnify the very societal stigmas it has the power to dismantle, affecting individuals' mental health and recovery processes adversely (Matsumoto et al., 2021; McNeil, 2021).

The widespread stigma surrounding SUD requires urgent and innovative solutions. Leveraging technology and social media, we can develop empathetic, supportive interventions that fight against this stigma (Rahaman et al., 2023). While research has explored mental health conversations and public perceptions on social media (Robinson et al., 2019), there remains a significant gap in efforts to destigmatize language in these discussions. Addressing this gap is crucial for fostering a more understanding and supportive environment for those affected by SUD.

Our work explores this opportunity and examines how stigmatizing language manifests in online communities and what solutions can be applied for de-stigmatizing such narratives (Table 1). Our study focuses on two research questions:

- **RQ1:** How does stigmatizing language manifest in non-drug-related Reddit communities when discussing SUD, and what are the underlying factors that contribute to such expressions?
- **RQ2:** How can we leverage LLMs to effectively de-stigmatize language, and what factors influence the success of this process?

To address these research questions, we collected over 1.2 million posts from non-drug-related subreddits, identifying 3,207 posts containing stigmatizing language towards people who use substances (PWUS). Leveraging large language models (LLMs), we developed a framework to characterize stigma based on conceptualization of Link and Phelan (2001) (*labeling, stereotyping, separation, status loss, and discrimination*) and transform them into more empathetic versions, resulting in 1,649 de-stigmatized pairs. Our analysis showed that stimulants and cannabis were the most frequently mentioned substances, with stigma more generally being associated with interpersonal relationships and moral judgments. Human evaluations showed that our Informed + Stylized system using GPT-4 can reduce stigma while preserving the original tone and relevance. Automatic evaluations further confirmed that our approach effectively reduced stigma while maintaining the stylistic and psycholinguistic properties of the original posts.

Our work makes several key contributions: (1) public release of a unique dataset of labeled stigmatizing posts; (2) demonstration of frameworks for de-stigmatizing text; and (3) exploration of the linguistic characteristics of stigma expressions towards people who use substances (PWUS) online. Additionally, this study introduces innovative uses of LLMs for generating suggestions to mitigate potentially harmful language.

2 Related Work

2.1 Stigma and Language

Stigma, a complex social phenomenon, is deeply intertwined with language. The linguistic relativity principle, as described by Whorf (1956), suggests that language shapes our perception of reality, including the formation of stigmatizing views. In the context of substance use experiences (SUE) and SUD, stigma can manifest in multiple forms: *self-stigma*, often rooted in shame (Luoma et al., 2012); *public stigma*, negative attitudes and beliefs which lead to discrimination and social exclusion; *structural stigma*, which limits resources and opportunities, embedded in societal norms and institutional practices (Hatzenbuehler, 2016).

Building upon Goffman (2009)'s foundational work, Link and Phelan (2001) conceptualized stigma as the co-occurrence of labeling, stereotyping, separation, status loss, and discrimination. This framework highlights how stigma operates alongside power inequalities, influencing both the individual and society at large. Research has explored the manifestation of stigma in online communities (Nippert et al., 2021), particularly within social media platforms (Clark et al., 2021), revealing both the potential for support and the amplification of existing stigmas, particularly among mental health and opiate-dedicated online communities (Chen et al., 2022; Eschliman et al., 2024).

Linguistic analysis has proven valuable in identifying and characterizing stigmatizing language. Dehumanizing labels and biased language can perpetuate negative stereotypes and contribute to discrimination (Giorgi et al., 2023). A recent study by the CDC found that while stigmatizing language in traditional media has decreased over time, its use on social media platforms has increased (McLaren et al., 2023), highlighting the need for targeted interventions in these spaces. The specific linguistic cues that distinguish stigmatizing content can differ between those with lived experience of substance use and those without, particularly regarding language considered "othering" and the use of labels like "addict" (Giorgi et al., 2023).

2.2 LLMs and Social Impact

LLMs have shown promise in addressing social issues like hate speech detection (Guo et al., 2023a) and bias mitigation (Schlicht et al., 2024). Recent research demonstrates that LLMs can perform on par with or even surpass benchmark machine learning models in identifying hate speech (Kumarage et al., 2024). Moreover, carefully crafted prompting strategies can leverage the knowledge encoded in LLMs to improve the detection of nuanced and context-dependent forms of hate speech (Guo et al., 2023b). However, the application of LLMs in sensitive domains raises ethical concerns. The "black box" nature of these models can make it difficult to understand their decision-making processes, raising issues of transparency and accountability (Guo et al., 2024). Additionally, biases in training data can be inadvertently perpetuated, leading to discriminatory outcomes (Mei et al., 2023). Addressing these ethical considerations is important for the responsible and equitable use of LLMs in de-stigmatization efforts.

2.3 De-stigmatization Efforts

Language-based interventions, such as the use of person-first language and empathetic communication, have shown promise in reducing stigma related to substance use. Research has demonstrated the impact of specific word choices on perceptions of individuals with SUD (Kelly et al., 2010). (McGinty et al., 2018) proposed a set of communication strategies to reduce stigma, including the use of sympathetic narratives, removing blame, and highlighting structural barriers to treatment. These findings contributed notably as the National Institute on Drug Abuse (NIDA) has also published guidelines for using non-stigmatizing language in discussions of SUD (NIDA, 2023).

AI-mediated interventions, particularly those leveraging LLMs, have the potential to scale and automate de-stigmatization efforts. While prior work has focused on text detoxification and bias reduction, in general, (Dale et al., 2021b; Mendelsohn et al., 2020; Pryzant et al., 2020), the specific application to SUD-related stigma remains underexplored. Additionally, (Spata et al., 2024) highlights the importance of using appropriate and well-validated measures to assess the effectiveness of interventions aimed at reducing stigma.

Our work builds upon the previous work by introducing a comprehensive computational approach to identify and categorize stigma. Focusing on public stigma, which we refer to as *directed stigma* throughout the paper, we operationalize Link and Phelan (2001)'s framework, analyzing instances of labeling, stereotyping, separation, and discrimination towards PWUS in discussions in non-drug-related Reddit communities .

3 Data

To achieve the study's objective of addressing stigmatizing language, we specifically focused on non-drug-related subreddits. This choice was made to capture how stigmatizing language manifests externally rather than within communities where members discuss their own experiences with drug use. Within these communities, stigmatizing language is often directed towards oneself (e.g., "No one should hire a junkie like me, I'm useless") or describes situations where members felt stigmatized (e.g., "My co-workers stopped having lunch with me when they learned I've been to rehab twice") which differs from the external stigmatizing language we aim to address. By focusing on non-drug-related subreddits, we ensure that our analysis targets the perpetuation of harmful stereotypes by those outside the drug-using community. This methodological choice is informed by the need to differentiate between internal and external stigma, as highlighted in the literature on stigma (e.g., Link and Phelan (2001)'s attributes of stigma).

Data Collection. To investigate the manifestation of stigmatizing language in non-drug-related online communities, we collected data from four popular subreddits: *r/unpopularopinion*, *r/offmychest*, *r/medicine*, and *r/nursing*. The first two subreddits were chosen for their high activity levels, diverse user bases, and relevance to discussions of substance use and SUDs. Recent research has highlighted the prevalence of stigmatizing language within medical professional communities as well on platforms such as Twitter, although the overall use of stigmatizing and de-stigmatizing language was found to be low (Scott Graham et al., 2022). Given the critical role that healthcare professionals play in the lives of individuals with SUD, we included two of the most popular subreddits for healthcare professionals; *r/nursing* and *r/medicine*.

We collected a total of 3.8 million posts from these subreddits. Table 2 shows the number of posts per subreddit. To ensure data quality, we excluded posts that were removed, deleted, or associated with deleted accounts. Additionally, we filtered out posts where the combined title and body text were less than 10 words to focus on substantive discussions. This resulted in a final dataset of 1.51

Subreddit	# Subscribers	# Posts	Date Range
r/medicine	478K	116,702	05/2005 - 12/2022
r/nursing	715K	212,755	12/2009 - 12/2022
r/offmychest	3.2M	1,607,341	02/2010 - 12/2022
r/unpopularopinion	4.3M	2,044,463	08/2013 - 12/2022

Table 2: Selected subreddits and raw #posts

million posts for analysis.

4 Methodology

To develop a stigma detection model and destigmatize texts, we first need to filter posts related to substance use. This is followed by detection and de-stigmatization processes. Figure 1 shows our study's overall pipeline. Each step is detailed in the following sections.

4.1 Developing a Stigma Detection Model

4.1.1 Filtering Substance Use-Related Posts

To identify posts containing stigmatizing language related to substance use, we first filtered posts collected from non-drug-related subreddits to find relevant discussions. Drug-related content includes any mention of illicit drugs or drug use (e.g., heroin, cocaine, LSD), prescription drugs that can be abused (e.g., narcotics, benzodiazepines), and other drugs that are not prescription but are also commonly abused (e.g., inhalants, bath salts). We began by manually annotating a random sample of 200 posts to establish a ground truth for relevance. Two annotators independently assessed each post, achieving 100% agreement on the presence or absence of substance use-related content.

Given the nuanced nature of language around substance use, including slang and idiomatic expressions, we used LLMs with few-shot prompting to identify posts within the larger dataset. Based on a comprehensive assessment of performance metrics, including precision, recall, F1-score, and estimate time (see Appendix A), we selected GPT-3.5 Turbo as the most suitable model for this task. As a result of Task 1, we identified around 33,064 posts containing at least one mention of drugs or drug-related content.

Validation Layer. Given the tendency of GPT-3.5 to overgeneralize, we implemented a validation layer using GPT-4 Turbo to re-evaluate all posts initially flagged as containing substance use-related content (N = 33,064). To evaluate the effectiveness of this validation layer, we randomly sampled 725 posts from the GPT-3.5 output (252 labeled as drug-related (*D*) and 473 as non-drug-related (*ND*)) and conducted a manual evaluation. The posts labeled as *D* by GPT-3.5 were then passed through the GPT-4 validation layer. Out of the 252 posts initially labeled as *D*, 212 were confirmed as *D* by GPT-4, resulting in an accuracy of $F1 = 0.86$. From the 33,064 posts labeled as *D* by GPT-3.5, 16,277 were validated as *D* by GPT-4.

4.1.2 Extracting Stigmatizing Language

The posts labeled as containing drug content were then labeled for their inclusion of stigmatizing language. Stigmatizing language could be in the form of directed language towards PWUS that perpetuates harmful stereotypes, expressions of internalized stigma (i.e., self-stigma), or illustrations of structural or systemic stigma (e.g., criminal justice towards PWUS in the United States). To do this, we took a random sample of 200 posts from the 16,277 posts labeled *D* and manually annotated for the inclusion of stigmatizing language. Any posts that contained directed stigmatizing language were also broken down into four attributes: 1) labeling, 2) stereotyping, 3) loss of power, and 4) discrimination. This process was re-iterated several times until substantial agreement was met ($k = 0.67$). The remaining posts were then labeled using GPT-4 Turbo using the prompt in Appendix B.

Explainability of Stigma Detection. In the pursuit of transparency and interpretability, we incorporated an explanation layer into our stigma detection model. Specifically, when the model identified a post as containing directed stigma towards PWUS, it was prompted to provide a detailed explanation for its classification by identifying the specific instances within the text that corresponded to each of the four elements of stigma outlined by Link and Phelan (2001): labeling, stereotyping, separation, and discrimination, mimicking our annotation process.

4.2 De-Stigmatizing Problematic Language

To address and mitigate the impact of stigmatizing language in texts, we used two different LLMs across three different Models. Our objective is to determine which model is most effective at transforming stigmatizing language into expressions that are more empathetic and inclusive.

Model 1: Baseline. In the baseline phase, we explored the capabilities of two LLMs in zero-shot de-stigmatization: GPT-4 Turbo and Llama 3-70B-Instruct. We provided the models with the original

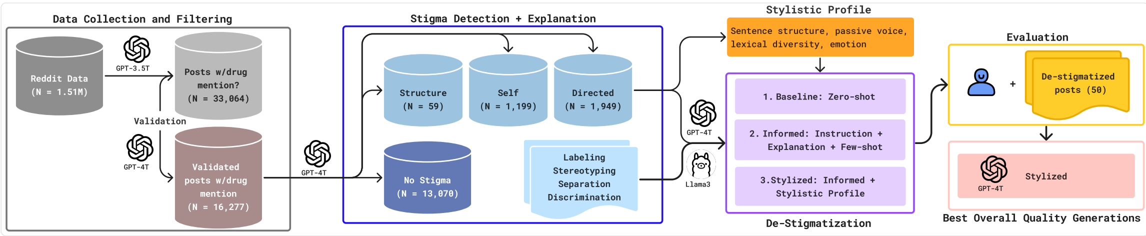

Figure 1: Full de-stigmatization pipeline.

	Stigma Type			
Substance Category	**Directed**	**Self**	**Structural**	**Total**
Stimulants	818	380	20	1218
Cannabis	515	276	27	818
Narcotics	501	250	18	769
Depressants	92	102	6	200
Hallucinogens	90	68	4	162
Reversal Agents	38	3	0	41
Drugs of Concern	7	7	0	14
Synthetic Cannabinoids	11	3	0	14
Other	4	3	1	8
Designer Drugs	6	0	0	6
Unspecified	537	475	9	1021

Table 3: Cross-tabulation of substance categories mentioned in a post by the type of stigmatizing language used. Note that multiple substance categories may be mentioned in the same post.

stigmatizing post and instructed them to generate a de-stigmatized version without any additional context or guidance. This approach allowed us to assess the inherent de-stigmatization capabilities of these models in the absence of explicit knowledge or stylistic refinements.

Model 2: Informed LLM. Inspired by the principles of "Constitutional AI," we enhanced the LLM prompts in Phase 2 with explicit instructions, definitions, and explanations related to stigma. Constitutional AI refers to the development and operation of AI models that adhere to the principles and legal standards, ensuring respect for human rights, ethical guidelines, and public accountability. Drawing upon the insights gained from our analysis of stigmatizing language (RQ1), we provided the model with a structured understanding of the four stigma elements (labeling, stereotyping, separation, and discrimination) and their manifestations in the context of substance use.

- **Labeling**: The model was instructed to identify and reword any labeling instances in the post, guided by a definition, explanation, and examples from RQ1 analysis.
- **Stereotyping, Separation, and Discrimination:** The model was tasked with addressing these three interrelated elements of stigma simultaneously. The prompt included definitions for each element, examples from RQ1 analysis, and an explanation as to why these elements are harmful to guide the LLM to mitigate these forms of stigma through rephrasing, reframing, or adding context.

By incorporating these explicit instructions and structured explanation of stigma, we aimed to guide the LLM in generating de-stigmatized outputs that actively addressed each of the four stigma elements identified in the original post.

Model 3: Informed LLM + Stylistic Considerations. Building upon the informed LLM approach of Phase 2, we further refined the de-stigmatization process by incorporating stylistic considerations. We aimed to ensure that the de-stigmatized output not only addressed the harmful content but also maintained the original post's emotional tone and stylistic features. To achieve this, we employed a combination of techniques:

- **Emotion Analysis:** We used a pre-trained, RoBERTa (Liu et al., 2019) model fine-tuned on the GoEmotions dataset (Demszky et al., 2020) [1], to classify the emotional tone of the original post and instructed the LLM to preserve this tone in the de-stigmatized version.
- **Punctuation and Syntax:** We analyzed the use of punctuation and sentence structure (i.e. sentence length variation) in the original post and encouraged the LLM to replicate these patterns in the output.
- **Stylistic Elements:** Posts were analyzed for phrase style, specifically the measure of textual lexical diversity (MTLD) (McCarthy and Jarvis, 2010) and the use of passive voice, to ensure that the de-stigmatized output maintained the original post's overall writing style.

These elements, plus the explanations, were used to produce de-stigmatized outputs that were less harmful and stylistically congruent with the original post, thereby maintaining the author's voice

[1] `https://huggingface.co/SamLowe/roberta-base-go_emotions`

and reducing the potential for inauthenticity.

4.2.1 Evaluation of De-Stigmatized Posts

Human Evaluation. To assess the effectiveness of our six systems (baseline, informed, and informed + stylized for GPT-4 and Llama3), we conducted a human evaluation with five reviewers on a random sample of 110 posts (a total of 660 generated texts). Our reviewers come from a variety of backgrounds, including HCI, NLP, and Social Computing. To evaluate the systems, we instructed our reviewers to analyze the generated text from each model and rank the models based on the overall quality, the extent of de-stigmatization, and the faithfulness of the outputs. Following traditional NLG assessments, quality was evaluated on criteria including naturalness, cohesion, human-likeness, and overall coherence (Howcroft et al., 2020). The assessment of de-stigmatization was judged based on removing negative or harmful stereotypes, and the systems with the least amount of labeling, stereotyping, separation, status loss, and discrimination. Faithfulness was evaluated based on the amount of transferred information from the original post without unnecessary details (Sai et al., 2022). Comprehensive evaluation guideline is provided in Appendix D.

Automatic Evaluation. To further evaluate the stylistic similarity between original posts and their de-stigmatized counterparts generated by our models, we conducted a linguistic analysis using LIWC (Boyd et al., 2022). We then performed a t-test to compare the linguistic features identified in both the original and de-stigmatized texts. Given the unique nature of our task, traditional metrics such as BLEU (Papineni et al., 2002) or ROUGE (Lin, 2004) were deemed unsuitable because the generated text and its original counterparts differ significantly in meaning. Additionally, the absence of pre-existing de-stigmatized versions of these texts prevented us from conducting comparative analyses with an established benchmark.

5 Experimental Results & Analysis

5.1 Characteristics of Stigmatizing Language

Mentioned Substances. Out of 16,277 posts discussing drugs, our stigma detection pipeline resulted in 3,207 posts containing stigmatizing language (Figure 1). Of these, 1,949 posts contained directed stigma, 59 represented systemic/structural stigma and 1,199 contained self-stigmatizing language. As shown in Table 3, analysis of stigmatizing posts revealed that stimulants like "meth" (methamphetamine) and "coke" (cocaine) were the most frequently mentioned drug categories, followed by cannabis ("weed", "pot") for all types of stigma. Posts that mentioned drug use terms like "drugs", "high", or "pills," but no specific substance were categorized as "Unspecified."

Anatomy of Stigma. To further understand *who*, did *what*, and *why* in the context of stigma towards PWUS in online discussions, we examined representative entities, subject-verb pairs, and topic models. Representative entities and subject-verb pairs reveal the *direction* of the mentions (who), while entity and substance frequencies highlight the targets of stigma (what). Topic modeling allows us to infer the underlying motivations and contexts of stigmatizing language (why). For this purpose, we used a multifaceted linguistic analysis: we first extracted subject-verb pairs using part of speech tagging in *spaCy* (Honnibal and Montani, 2020), classified emotions toward these pairs in each post using GoEmotions (Demszky et al., 2020) and RoBERTa (Liu et al., 2019), and performed topic modeling with BERTopic (Grootendorst, 2022) and KeyBERT (Grootendorst, 2020).

Within the posts showing directed stigma (Appendix C), we primarily observe expressions of *sadness* and *annoyance*, with some *neutrality*. Notably, interpersonal relationships surface as a key theme, featuring mentions of family members like "sister," "dad," and "mother" alongside substances like "cannabis" and "amphetamines." This aligns well with the overall prevalence of stimulants and cannabis in substance mentions (Table 3). The dominant topic, "Cannabis and Legalization Stigma" centers on these substances, often referred to as "it," in a *neutral* tone primarily related to "smoking." Following closely is "Stigma Toward Interpersonal Relationships," characterized by expressions of knowledge (*I know*) from the subject *"I"* directed towards family members, often tinged with *sadness*. Another notable topic, "Moral Judgments of Others," reveals *annoyance* (*I hate*) towards individuals like "neighbors," "homeless," and "junkies" associated with "heroin" and other drugs.

Shifting to self-stigmatizing posts, we find distinct emotional undertones and actions. While interpersonal entities are less prominent compared to directed stigma, these posts feature more action verbs and a wider variety of substances. The primary topic, "Depression around Sobriety," is marked by expressions of possession (*I have*) and state of being (*I am*) in relation to *depression*, *sobriety*, and *quitting*. Disturbingly, another topic reveals a *desire* for *overdose*, specifically mentioning narcotics like "fentanyl," "dilaudid," and "tramadol," alongside benzodiazepines like "Xanax" and "clonazepam," a combination known to be potentially lethal due to respiratory arrest.

Finally, while only a few posts showed structural stigma (N = 59), making it hard to analyze topics, the emotions in these posts were mainly neutral.

5.2 De-Stigmatization with LLMs

Human Evaluation. Human evaluation of 110 randomly selected and destigmatized posts is shown in Table 4. Our reviewers found the Informed + Stylized system with GPT-4 the best "Overall Quality" for de-stigmatizing posts and the "Most Faithful", striking a balance between reducing stigma and maintaining relevance and tone with the original message. Interestingly, the "Most De-stigmatized" system was the Baseline with GPT-4. While this system may be effective at removing overt stigmatizing language, the lack of contextual awareness and stylistic preservation resulted in posts sounding generic, insincere, unfaithful, or disconnected from the original poster's voice. For example, with the Baseline + GPT-4, this post:

"Decriminalizing drugs is a dangerous and dumb idea and shouldn't be done. what a wonderful idea, let's give druggies more room and less consequences for doing the hard stuff. when out, people need to keep an eye out to not accidentally step on used needles, crime is up from junkies who do anything to afford their next fix." is de-stigmatized to:

"Decriminalizing drugs is a complex and controversial topic that merits a thoughtful discussion. There are valid concerns about the potential increase in visibility and accessibility of substances.states that have adopted more lenient policies towards drug possession, see increased public drug use and challenges, such as discarded needles and crime, which worries local residents."

For practical applications, this could affect the model's ability to foster genuine empathy and understanding in online communities. Table 1 displays a successful de-stigmatized post using our best model. The revised post is less harmful and stereotypical but maintains the speaker's tone and overall message.

Automatic Evaluation. We conducted a stylistic similarity analysis using LIWC to compare original stigmatizing posts with their de-stigmatized versions generated by our top-rated system (Informed + Stylized GPT-4). A pairwise two-way t-test showed no significant differences in means across all LIWC variables between the two sets of posts. While certain categories like bigwords (use of six-letter words or more) and cogproc (cognitive processes) were more common in de-stigmatized posts, the overall psycholinguistic properties were largely maintained. This result is promising as it shows our de-stigmatization approach effectively reduced stigma while preserving the original style and emotional tone, essential for authenticity.

6 Discussion

Stigma also stems from personal connections. Our findings showed a complex landscape of stigma within non-drug-related online communities where discussions about substance use often become entangled with interpersonal relationships and ingrained societal biases - particularly towards specific substances, namely stimulants (e.g., methamphetamine) and cannabis (e.g., "weed," "pot"). The frequent mentions of these substances within a stigmatizing context may reflect societal concerns about their visibility and impact, aligning with our topic modeling results, where the dominant topic in directed stigma is "Cannabis Legalization Stigma." These findings highlight the role of close relationships (family, friends) in both expressing and experiencing stigma. For instance, within the topic "Interpersonal Stigma," we observe individuals expressing sadness and using the verb "know" when discussing family members struggling with substance use. This underscores the need for de-stigmatization efforts to extend beyond public forums and into private spheres, as stigma from close social circles can be particularly harmful due to the emotional weight and potential for isolation (Luoma et al., 2012).

The online nature of these interactions presents a duality of stigma manifestations that is important to understand when developing any intervention. While anonymity might offer a shield for individuals to express stigmatizing views they might suppress offline, it could also create a space for open dialogue and support. The disinhibition afforded by online platforms could lead to more candid discussions about SUD, potentially challenging stigma

Model	LLM	Best Overall Quality	Most De-Stigmatized	Most Faithful
Informed + Stylized	GPT4	**37**	18	**49**
Informed	GPT4	24	7	33
Informed	Llama	19	8	16
Informed + Stylized	Llama	13	3	6
Baseline	Llama	9	32	2
Baseline	GPT4	6	**40**	2

Table 4: Frequency of evaluation metrics by systems for 110 de-stigmatized posts.

through shared experiences and mutual understanding. However, it may also create a space for misinformed judgments and harmful stereotypes, as anonymity can reduce accountability.

When considering de-stigmatization efforts, any digital intervention should consider the social actors in addition to the social constructs (e.g. hospitals, employers). This would be considerably important in collectivist communities (e.g. Indian or Middle Eastern) where stigma towards family members with an SUD (i.e. *affiliate stigma*) may prevent families from providing the necessary medical support to their loved ones and ultimately delaying treatment (Corrigan et al., 2006).

LLMs can be guided by explanation and stylistic information. In our de-stigmatization efforts, we intentionally avoided providing the LLMs with a rigid definition of "de-stigmatized." Instead, we adopted a more nuanced approach, drawing inspiration from the principles of "Constitutional AI" and prior work on text detoxification and bias reduction using LLMs (Dale et al., 2021a; Mendelsohn et al., 2021; Pryzant et al., 2020). We focused on explaining why specific phrases might be problematic and instructed the model to address these issues, constitutionally, while preserving the original style. For instance, to tackle separation, the LLMs were guided to draw equivalences between individuals with SUD and those without, emphasizing shared humanity. Labeling was addressed by replacing derogatory terms like "junkie" with person-centered language like "person with a substance use disorder," mitigating the over-generalization tendencies of LLMs. Stereotyping and discrimination were handled by re-framing generalizations and removing any implications of discrimination, promoting a more empathetic understanding of individuals struggling with SUD.

Most de-stigmatized does not mean most pragmatic. While the baseline model removes stigmatizing language, it often does so at the expense of nuance and context. For instance, evaluators noted that the baseline model sometimes "terribly misunderstood the post," resulting in generic or insincere responses that failed to capture the original poster's intent. This highlights the importance of removing stigma and preserving the authenticity and emotional tone of the original message. Our findings emphasize the importance of striking a balance between promoting empathetic language and providing overly refined language, which might trivialize the experiences of individuals with SUD or avoid addressing the root causes of stigma.

7 Conclusion

This study investigated the manifestations of stigma towards PWUS in four popular non-drug-related subreddits (*r/unpopularopinion*, *r/offmychest*, *r/nursing*, *r/medicine*). We identified 3,207 posts containing one of three main types of stigma (self, structural, and directed). Given the contextual nuance of self and structural stigma, we focused our efforts on de-stigmatizing instances of directed stigma (N = 1,649). Experimenting with three different models and two different LLMs (GPT-4 and Llama), the model that used the conceptualization of stigma (Link and Phelan, 2001), few-shot examples, and the original post's stylistic profile generated the most faithful and appropriate destigmatized texts. Our exploration of LLM-based de-stigmatization demonstrates the potential of these models to transform harmful language into more empathetic expressions while emphasizing the importance of preserving authenticity and the original poster's voice. While our focus has been on SUD stigma, the insights and methodologies presented here have broader implications for understanding and addressing stigma related to other marginalized groups. Future work could explore the role of misinformation in perpetuating stigma and leverage external knowledge bases (e.g. DrugBank) to develop more informed and effective de-stigmatization strategies. By integrating these approaches, we can create a more supportive and inclusive online environment for individuals affected by stigma, ultimately promoting understanding, empathy, and recovery.

8 Limitations

Our findings primarily apply to English-speaking populations on one specific social media platform, which may not be generalizable to other linguistic or cultural contexts. We selected certain subreddits based on our assessment of relevance, which may have limited the breadth of our data; exploring additional subreddits could potentially provide a more comprehensive view. The performance and accuracy of the models we used, dependent on their training data, may not capture all nuances of stigmatizing language. Despite our ethical considerations, the automated analysis of sensitive topics like SUD carries risks of misinterpretation, necessitating ongoing research and continuous evaluation of ethical challenges in using large language models.

9 Ethics Statement

We acknowledge the diversity of perspectives on substance use and advocate for harm reduction strategies. All data was publicly available at the time of collection, and no direct interaction occurred between researchers and users. Our research was exempt from review by our institution's Internal Review Board (IRB). We adhere to strict data protection measures and have slightly altered any quotes to preserve anonymity and post integrity. Our goal is not to erase personal experiences but to reframe them in less harmful ways, aligned with the original sentiment. The discussions in this paper should not be interpreted to suggest anyone's lived experience is more valid than another.

References

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.

Centers for Disease Control and Prevention. 2023. Reducing stigma to prevent opioid overdose. `https://www.cdc.gov/stop-overdose/stigma-reduction/index.html`. Accessed: 2024-06-15.

Annie T. Chen, Shana Johnny, and Mike Conway. 2022. Examining stigma relating to substance use and contextual factors in social media discussions. *Drug and Alcohol Dependence Reports*, 3:100061.

Olivia Clark, Matthew M Lee, Muksha Luxmi Jingree, Erin O'Dwyer, Yiyang Yue, Abrania Marrero, Martha Tamez, Shilpa N Bhupathiraju, and Josiemer Mattei. 2021. Weight stigma and social media: evidence and public health solutions. *Frontiers in nutrition*, 8:739056.

Patrick W Corrigan, Amy C Watson, and Frederick E Miller. 2006. Blame, shame, and contamination: the impact of mental illness and drug dependence stigma on family members. *Journal of family psychology*, 20(2):239.

David Dale, Igor Markov, Varvara Logacheva, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021a. SkoltechNLP at SemEval-2021 task 5: Leveraging sentence-level pre-training for toxic span detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 927–934, Online. Association for Computational Linguistics.

David Dale, Anton Voronov, Daryna Dementieva, Varvara Logacheva, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021b. Text detoxification using large pre-trained neural models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7979–7996, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. ArXiv:2005.00547 [cs].

E.L. Eschliman, K. Choe, A. DeLucia, E. Addison, V.W. Jackson, S.M. Murray, D. German, B.L. Genberg, and M.R. Kaufman. 2024. First-hand accounts of structural stigma toward people who use opioids on Reddit. *Social Science and Medicine*, 347.

Salvatore Giorgi, Douglas Bellew, Daniel Roy Sadek Habib, Garrick Sherman, João Sedoc, Chase Smitterberg, Amanda Devoto, McKenzie Himelein-Wachowiak, and Brenda Curtis. 2023. Lived experience matters: automatic detection of stigma on social media toward people who use substances. *arXiv preprint arXiv:2302.02064*.

Erving Goffman. 2009. *Stigma: Notes on the management of spoiled identity*. Simon and schuster.

Maarten Grootendorst. 2020. Keybert: Minimal keyword extraction with bert.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Keyan Guo, Alexander Hu, Jaden Mu, Ziheng Shi, Ziming Zhao, Nishant Vishwamitra, and Hongxin Hu. 2023a. An investigation of large language models for real-world hate speech detection. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1568–1573.

Keyan Guo, Alexander Hu, Jaden Mu, Ziheng Shi, Ziming Zhao, Nishant Vishwamitra, and Hongxin Hu. 2023b. An investigation of large language models for real-world hate speech detection. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1568–1573.

Zhijun Guo, Alvina Lai, Johan Hilge Thygesen, Joseph Farrington, Thomas Keen, and Kezhi Li. 2024. Large language model for mental health: A systematic review. *arXiv preprint arXiv:2403.15401*.

Mark L. Hatzenbuehler. 2016. Structural stigma: Research evidence and implications for psychological science. *American Psychologist*, 71(8):742–751. Place: US Publisher: American Psychological Association.

Matthew Honnibal and Ines Montani. 2020. spacy: Industrial-strength natural language processing in python.

David M Howcroft, Anya Belz, Miruna Clinciu, Dimitra Gkatzia, Sadid A Hasan, Saad Mahamood, Simon Mille, Emiel Van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: Nlg needs evaluation sheets and standardised definitions. In *13th International Conference on Natural Language Generation 2020*, pages 169–182. Association for Computational Linguistics.

John F. Kelly, Sarah J. Dow, and Cara Westerhoff. 2010. Does our choice of substance-related terms influence perceptions of treatment need? an empirical investigation with two commonly used terms. *Journal of Drug Issues*, 40(4):805–818.

Tharindu Kumarage, Amrita Bhattacharjee, and Joshua Garland. 2024. Harnessing artificial intelligence to combat online hate: Exploring the challenges and opportunities of large language models in hate speech detection. *arXiv preprint arXiv:2403.08035*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Bruce G. Link and Jo C. Phelan. 2001. Conceptualizing Stigma. *Annual Review of Sociology*, 27(1):363–385. _eprint: https://doi.org/10.1146/annurev.soc.27.1.363.

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].

Jason B. Luoma, Barbara S. Kohlenberg, Steven C. Hayes, and Lindsay Fletcher. 2012. Slow and steady wins the race: A randomized clinical trial of acceptance and commitment therapy targeting shame in substance use disorders. *Journal of Consulting and Clinical Psychology*, 80(1):43–53. Publisher: American Psychological Association.

Atsushi Matsumoto, Claudia Santelices, and Alisa K Lincoln. 2021. Perceived stigma, discrimination and mental health among women in publicly funded substance abuse treatment. *Stigma and Health*, 6(2):151.

Philip M McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Emma McGinty, Bernice Pescosolido, Alene Kennedy-Hendricks, and Colleen L. Barry. 2018. Communication strategies to counter stigma and improve mental illness and substance use disorder policy. *Psychiatric Services*, 69(2):136–146.

Nilay McLaren, Christopher M. Jones, Rita Noonan, Nimi Idaikkadar, and Steven A. Sumner. 2023. Trends in stigmatizing language about addiction: A longitudinal analysis of multiple public communication channels. *Drug and Alcohol Dependence*, 245:109807.

Sandra R McNeil. 2021. Understanding substance use stigma. *Journal of Social Work Practice in the Addictions*, 21(1):83–96.

Katelyn Mei, Sonia Fereidooni, and Aylin Caliskan. 2023. Bias against 93 stigmatized groups in masked language models and downstream sentiment classification tasks. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1699–1710, New York, NY, USA. Association for Computing Machinery.

Julia Mendelsohn, Ceren Budak, and David Jurgens. 2021. Modeling framing in immigration discourse on social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2219–2263, Online. Association for Computational Linguistics.

Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. 2020. A framework for the computational linguistic analysis of dehumanization. *Frontiers in artificial intelligence*, 3:55.

NIDA. 2023. Words matter - terms to use and avoid when talking about addiction. https://nida.nih.gov/research-topics/addiction-science/words-matter-preferred-language-talking-about-addiction. Accessed: 2024.

Kathryn E Nippert, A Janet Tomiyama, Stephanie M Smieszek, and Angela C Incollingo Rodriguez. 2021. The media as a source of weight stigma for pregnant and postpartum women. *Obesity*, 29(1):226–232.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the*

40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. In *Proceedings of the aaai conference on artificial intelligence*, volume 34, pages 480–489.

Md Saidur Rahaman, MM Ahsan, Nishath Anjum, Md Mizanur Rahman, and Md Nafizur Rahman. 2023. The ai race is on! google's bard and openai's chatgpt head to head: an opinion article. *Mizanur and Rahman, Md Nafizur, The AI Race is on*.

Patrick Robinson, Daniel Turk, Sagar Jilka, and Matteo Cella. 2019. Measuring attitudes towards mental health using social media: investigating stigma and trivialisation. *Social psychiatry and psychiatric epidemiology*, 54:51–58.

Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Computing Surveys (CSUR)*, 55(2):1–39.

Ipek Baris Schlicht, Defne Altiok, Maryanne Taouk, and Lucie Flek. 2024. Pitfalls of conversational llms on news debiasing. *arXiv preprint arXiv:2404.06488*.

Georg Schomerus, Michael Lucht, Anita Holzinger, Herbert Matschinger, Mauro G Carta, and Matthias C Angermeyer. 2011. The stigma of alcohol dependence compared with other mental disorders: a review of population studies. *Alcohol and alcoholism*, 46(2):105–112.

S. Scott Graham, F.N. Conway, R. Bottner, and K. Claborn. 2022. Opioid use stigmatization and destigmatization in health professional social media. *Addiction Research and Theory*.

Angelica Spata, Ishita Gupta, M Kati Lear, Karsten Lunze, and Jason B Luoma. 2024. Substance use stigma: A systematic review of measures and their psychometric properties. *Drug and Alcohol Dependence Reports*, page 100237.

Substance Abuse and Mental Health Services Administration. 2023. Samhsa announces nsduh results detailing mental illness and substance use levels in 2021. https://www.samhsa.gov/newsroom/press-announcements/20230104/samhsa-announces-nsduh-results-detailing-mental-illness-substance-use-levels-2021.

Benjamin Lee Whorf. 1956. *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf*. MIT Press.

A Comparison of LLMs for Labeling Drug Mention

We examined various LLMs (combination of open-source and proprietary) to differentiate between drug-related and non-drug-related posts on Reddit, using a dataset of 200 manually annotated posts. To assess the performance of each model, we calculated the F-1 score, which is a measure of a test's accuracy that considers both precision and recall. Additionally, we analyzed the total time and cost required to process this amount of posts. These findings are detailed in the table provided in Table 5. This table helps to illustrate not only the effectiveness of each model in terms of accuracy but also their efficiency and economic viability for similar tasks.

Model	F1	Total Time	Cost (USD)	RPM
GPT 3.5-Turbo	0.78	9.52 s	0.07	3,500*
GPT 4-Turbo	0.9	19.05 s	1.31	500*
Mistral	0.48	330.60 s	0	300**
Llama3-8B	0.38	59.9 s	0	600***

Table 5: Comparison on four LLMs considered to label 1.51M posts for the mention of drugs or drug use based on a random sample of 200 manually-annotated posts.
'*' based on OpenAI Tier 3 usage (see https://platform.openai.com/docs/guides/rate-limits/usage-tiers?context=tier-three)
'**' based on Hugging Face Inference API rate limit per hour
'***' based on Together.ai API rate per second for Paid Tier (https://docs.together.ai/docs/rate-limits).

B Prompts

In our study, we implemented a multi-step pipeline using different prompts for each stage, which includes data filtering, stigma detection with explanations, and destigmatization. The prompts tailored for data filtering, stigma detection, and destigmatization are detailed in Figures 2, 3 and 4. This structured approach ensures efficient handling and analysis of stigmatizing content in social media posts.

C Data Analysis

In our study, we conducted a comprehensive linguistic analysis of online posts about drug use and addiction-related stigmas. We extracted and analyzed representative entities, subject-verb pairs, Instruction

```
Objective: Identify references to drugs or people who use drugs in each post.
Include:
• Illicit Drugs: All controlled substances with no legal usage such as cannabis, heroin, cocaine, methamphetamine
• Prescription Drugs: Drugs that are often abused even if they have legitimate medical uses such as opioids,
  benzodiazepines
• Other Drugs: Non-prescription drugs known to be abused such as inhalants, k2, bath salts
• Explicit mentions of drug use, abuse, or addiction related terms such as "getting high", "stoned"
Exclude:
• Tobacco, nicotine, or alcohol unless explicitly linked to drug use
• Do not include medical or psychological discussions unless there is a direct and clear mention of drug use or abuse
Clarifications:
• Mental health discussions should not be labeled as 'D' unless there is an explicit mention of drugs as defined above
• Use 'ND' for posts that discuss health or psychological issues without specific drug references
Language Cues:
• Focus on clear drug-related terminology such as "junkie", "addict" and slang
• If a post is ambiguous and does not clearly fit the drug reference criteria, label as 'ND'
Response Requirement:
• Respond with either 'D' (Drug) or 'ND' (Non-Drug) based on these guidelines. No additional commentary is needed.
```

Few-Shot Examples

```
###POST: "I hope my junkie sister OD's or disappears out of our lives My sister is an alcoholic junkie who has 2 DUIs under
her belt as well as loves taking Xanax and alcohol together and wreaking havoc for our family and even strangers."
###LABEL: D
###POST: "My drinking problem is way worse than anyone knows. I'm an alcoholic and the people around me know that, because
I told them. But if I hadn't, they wouldn't have guessed, judging by their reactions of shock and disbelief and the actual
,are you sure you're not exaggerating, that I got when I came clean a couple years ago."
###LABEL: ND
###POST: "I cut tonight I cut myself tonight after I've been clean for a few months now and I'm honestly just broken
inside. I had a huge fight with my girlfriend, I'm worried we might break up. I just kinda lost control, grabbed my
razorblade and cut. I feel like crap and am honestly just lost right now."
###LABEL: ND
###POST: "Hate my new job I just got hired at Taco Bell a few weeks ago and I really hate it. The only thing that's gets me
through it is smoking weed. I wonder if i should just quit or no call no show?"
###LABEL: D
...
```

Target

```
###POST: Left a best friend hanging. Let one guy at the beach take a fall for all the weed and alcohol. I put in to smoke
and all 15-18 of us just let him take all of the blame.
###LABEL:
```

Figure 2: Few-shot prompting to determine whether a post contains a reference to illicit substance use.

and utilized topic modeling to identify themes in direct and self-stigmatizing posts. These topics were organized by names, representative keywords, dominant emotions, and frequent verb-subject pairs, presented in Tables 6 and 7. For linguistic processing, we used *spaCy* for subject-verb extraction, GoEmotions and RoBERTa for emotion classification, and BERTopic and KeyBERT for topic modeling. This multi-dimensional approach provided a detailed understanding of the discourse within these posts.

D Human Evaluation

We provided the following instructions to guide the evaluation of de-stigmatized texts. These guidelines were shared with our reviewers together with the generated texts from the six systems and forms for ranking the models. Each reviewer judged 20 to 30 posts independently.

Instructions: Please read the original post and the generated posts from each of the six systems carefully. For each of the following questions, select the system that best represents your evaluation. Use the space provided for any additional comments you may have.

Overall Quality:

Assess the overall quality of the generated text with respect to the following measures in addition to de-stigmatization and faithfulness:

- **Naturalness:** The degree to which the output is likely to be used/chosen by a native speaker in the given context/situation.
- **Cohesion:** The text should be a well-organized and coherent body of information, not just a dump of related information. Specifically, the sentences should be connected to one another, maintaining good information flow and logic.
- **Appropriateness:** The degree to which the output as a whole is appropriate in the given context/situation. E.g., “does the text appropriately consider the parents’ emotional state in the given scenario?”
- **Human-Likeness:** The degree to which the output could have been produced by a human.

Which system generated the text that with the best overall quality (content, form, de-stigma)?

Instruction

```
Instructions: You are an expert in identifying stigma related to drug use and addiction in social media posts. Your task is
to analyze each post and determine if it contains stigmatizing language.
Label:
 • If the post contains stigmatizing language, label it as "S" (Stigma).
 • If the post does not contain stigmatizing language, label it as "NS" (Non-Stigma).
Explanation:
 • For "S" (Stigma) labels:
   • If the post demonstrates stigmatizing language (directed at others), analyze it based on the following four attributes
     and provide an explanation for each attribute:
     • Labeling: Does the post use derogatory or othering language related to drug use/addiction?
     • Stereotyping: Does the post reinforce negative generalizations about people who use drugs?
     • Separation: Does the post create a divide between people who use drugs and those who don't?
     • Discrimination: Does the post imply or suggest unfair treatment based on drug use?
   • If the post demonstrates self-stigma or structural stigma, provide a brief, one-sentence explanation of the type of
     stigma present.
 • For "NS" (Non-Stigma) labels: No explanation is needed.
```

Definition

```
Definitions:
 • Stigma: Negative attitudes or beliefs directed at individuals based on their drug use.
 • Self-Stigma: Internalization of negative societal attitudes about drug use.
 • Structural Stigma: Societal systems, policies, or practices that disadvantage people who use drugs.
Additional Considerations:
 • Consider the context of the post and the author's intent.
 • Focus on language that is harmful, discriminatory, or perpetuates negative stereotypes.
Task: Read the following post, label it as "S" or "NS", and provide the appropriate explanation:
```

Few-Shot Examples

```
###POST: "Once an addict, always an addict. They'll never change."
###LABEL: "S, Labeling: 'addict', Stereotyping: Assumes that people with addiction are incapable of recovery or personal
growth, Separation: Creates a sense of hopelessness and perpetuates the idea that people with addiction are fundamentally
different, Discrimination: Discourages offering support or opportunities to individuals in recovery due to a belief they'll
relapse."
###POST: "He's just a junkie looking for his next fix."
###LABEL: "S, Labeling: 'junkie', a derogatory term, Stereotyping: Reduces the person's identity to their substance use,
ignoring other aspects of their life, Separation: Creates a distance between 'us' (non-users) and 'them' (people who use
substances), Discrimination: Implies the person is solely motivated by drug use and lacks other values or goals."
###POST: "She's constantly nodding off at work. Probably using again."
###LABEL: "S, Labeling: Implies the person has a substance use disorder by suggesting they are 'using again,' Stereotyping:
Generalizes that the person's behavior (nodding off) is solely due to drug use, Separation: Creates a distinction between
the person and their colleagues, implying they are unreliable or irresponsible due to assumed substance use, Discrimination:
Could lead to unfair treatment at work based on an assumption of relapse."
###POST: "I'm concerned about my brother's substance use. I'm learning about available resources to help him."
###LABEL: NS
...
```

Target

```
###POST: Left a best friend hanging. Let one guy at the beach take a fall for all the weed and alcohol. I put in to smoke
and all 15-18 of us just let him take all of the blame.
###LABEL:
```

Figure 3: Few-shot prompting to determine whether a post contains stigmatizing language towards PWUS.

De-stigmatization:

Which system's generated post is the most de-stigmatized?

Effectively De-stigmatized:

Which system's post has been the most effectively de-stigmatized, removing negative or harmful stereotypes? Remember stigma is defined as the co-occurring processes of labeling, stereotyping, separation, status loss, and discrimination. In the context of SUD, it can look something like this:

Labeling: involves identifying individuals as different based on certain characteristics—in this case, their substance use. For those with SUD, labels such as "addict" or "alcoholic" can be affixed.

Stereotyping: involves ascribing a fixed set of beliefs or characteristics to individuals based solely on their disorder.

Separation/Status Loss: the social distancing of a group perceived as different or undesirable. This separation is partly due to the fear and misunderstanding surrounding the disorder.

Discrimination: Discrimination can be both formal and informal, impacting various aspects of life, including employment and social interactions.

Faithfulness:

Evaluate whether the posts generated by each system contain all the required information from the original post without unnecessary details. Which system has the most faithful result?

General Feedback:

Please provide any general feedback or additional comments regarding your evaluation of the texts.

De-stigma with Explanation

```
if step Label:
    instruction = "Rewrite this post to remove any and all labeling."
    definition = "Labeling includes the use of derogatory or othering language related to drug use/addiction."
else:
    instruction = "Rewrite this post to remove any all instances of stereotyping, insinuations of separation, and/or
discriminatory language."
    definition = "Stereotyping reinforces negative generalizations about people who use drugs. Separation creates a divide
between people who use drugs and those who don't. Discrimination implies or suggests unfair treatment based on drug use."

prompt = f"""
    {instruction};
    {definition};
    Only rewrite the relevant parts of the post, do not rewrite the whole post. Do not change the meaning of the post or add
any new information.
    Example:
    post: "My mom is an addict"; This post uses the term 'addict'
    rewrite: "My mom has a substance use disorder"

    Do not include "Here is the rewritten post:" in your response. Just return the rewritten post.
    """
```

De-stigma with Explanation and Stylized

```
if step Label:
    instruction = "Rewrite this post to remove any and all labeling."
    definition = "Labeling includes the use of derogatory or othering language related to drug use/addiction."
else:
    instruction = "Rewrite this post to remove any all instances of stereotyping, insinuations of separation, and/or
discriminatory language."
    definition = "Stereotyping reinforces negative generalizations about people who use drugs. Separation creates a divide
between people who use drugs and those who don't. Discrimination implies or suggests unfair treatment based on drug use."

prompt = f"""
    {instruction};
    {definition};
    Only rewrite the relevant parts of the post, do not rewrite the whole post. Do not change the meaning of the post or add
any new information.
    Example:
    post: "My mom is an addict"; This post uses the term 'addict'
    rewrite: "My mom has a substance use disorder"

    Do not include "Here is the rewritten post:" in your response. Just return the rewritten post.
    """
```

Plus Style Profile:

```
{
    "role": "user",
    "content": post + ";" + ex + ";" + style_instruct
}
```

Figure 4: Few-shot prompting for de-stigmatizing language towards PWUS, explanation and explanation plus style profile.

Name	Representation	Top Emotion	Top Verb-Subject Pairs	Example
Cannabis Legalization Stigma	marijuana, cannabis, weed, drugs, addicts, sober, smoking, heroin, pot, smokers	neutral	{'it', 'is'}: 126, {'i', 'have'}: 117, {'i', 'know'}: 91	Your addiction and dependence isn't slighter than mines and vice versa. Just because weed doesn't have physiological symptoms of wd it doesn't mean it doesn't fuck up potheads who have to go without smoking for, say, week. Mind your own business.
Interpersonal Stigma	rehab, sister, family, dad, grandmother, parents, mother, father, drugs, mom	**sadness**	{'i', 'know'}: 381, {'i', 'have'}: 260, {'i', 'want'}: 256	I wish my sister would just go to prison and leave my family alone. About 10 years ago my sister got into a bad wreck. She was in a coma for a week and now has traumatic brain injury.
Moral Judgments on Addiction	homelessness, homeless, neighbor, neighbors, neighbour, junkies, neighborhood, drugs, heroin, cops	**annoyance**	{'i', 'see'}: 23, {'i', 'know'}: 21, **{'i', 'hate'}: 19**	This is completely ignoring the fact that drugs are the reason they are homeless in the first place. Some of the other comments were saying that they do drugs so why should they judge a homeless person doing drugs. This kind of justification seems insane to me. Just because you are ruining your life, doesn't mean that you should advocate for other people to ruin their lives. And I don't even want to get into the hundreds of drug subreddits like r/heroin, r/meth, and r/crack where people are posting about and bragging about their dangerous drug addictions.
Moral Judgements and Amphetamine Use	adderall, amphetamine, amphetamines, adhd, stimulant, prescriptions, prescription, drugs, medication, prescribed	neutral	{'i', 'have'}: 5, {'i', 'had'}: 4, **{'i', 'hate'}: 4**	I live in a college town and adderall/vyvanse use is insane. Some use it to study, some use it to party and some use it to game for days. All these people eventually can't operate without the pills. It leads to serious rage issues and mood swings. My roommate spends around $300/month on someone else's adderall. Here are some facts- he will exhaust you with hours of pointless stories and ramblings then get mad when you don't listen. He literally can't shut the hell up. Just like a tweaker.
Drug Use Consequences	vicodin, smoked, smoking, toxic, camping, run, thinking, scared, needle, crystal	neutral	**{'i', 'wanted'}: 6,** {'i', 'know'}: 5, **{'it', 'feels'}: 5**	Shot of meth feels like you've finally crossed that line you swore you'd never cross. You know the one-it looked impossibly far away back when you were naive enough to promise yourself you'd always stick to smoking. When you truly believed you would never allow yourself to become one of those needle freak losers.

Table 6: Summary of topics from direct stigmatizing posts. Interpersonal entities in blue, substances in green, and actions in purple.

Name	Representation	Top Emotion	Top Verb-Subject Pairs	Example
Sobriety & Family Struggles	depressed, depression, alcoholic, sober, stay, addiction, parents, drinking, quit, mother	**sadness**	{'(i, have)': 410, '(i, 'm)': 360, '(i, want)': 345}	I've been trying to come out of my isolation, they don't really care, and would rather keep my home and safe. so they screamed at me because I stayed out with my friends too late. I do not have that freedom anymore. I felt like I wanted to stay with my friends until I got comfortable. This was the first time I had hung out with them in a month, and I wasn't even enjoying it. I was uncomfortable. I tried weed, got even more uncomfortable. I can almost never turn down drugs. I am such a pathetic fucking junky.
Prescription Medication	adderall, medications, prescription, adhd, medication, opiate, prescribed, meds, pharmacy, xanax	**disappointment**	{'(i, have)': 78, '(i, feel)': 74, '(i, know)': 46}	I apologize if this doesn't make sense. I'm not very good at explaining things. I'm sure a lot of people will just judge me for being a whiny addict and say "well don't do drugs and you wouldn't even be in this situation, duh". I get it, most people think that all junkies should be "thrown on an island to die" and the world would be a much better place.
Overdose Death & Suicide Ideation	overdosed, xanax, fentanyl, dilaudid, acetaminophen, 600mg, 30mg, tramadol, prozac, clonazepam	**desire**	{'(i, want)': 21, '(i, 'm)': 15, '(i, know)': 9}	It didn't work, I'm not dead. I looked up what would happen if I took a shit ton of vyvanse and apparently seizures and heart failure are likely. shitty death but I needed to organize my stuff so it's easier to move or get rid of. so I took all the ones I had in the bottle. I spent literally the last couple hours writing suicide notes for nothing.
Struggles with Intimate Partners	youll, bye, alcoholic, leave, escaping, whisper, soul, leaving, pot, lifennim	**sadness**	{'(i, want)': 11, '(i, 'm)': 10, '(i, m)': 9}	Just an out of the blue rant from a worthless junkie.... don't bother. oh god I miss you so much. we know each other inside and out and have been through it all. I never thought you'd take me back ever from all the horrible shit I've done then to my surprise you took my back a second time even tho I ran away for months on end with no word or attempt of communication, getting high and drunk 24/7 and randomly showed up back home at 3 in the morning just to leave two days later and repeat my actions. then you moved a whole other province away to get back with me just for me to turn back to drugs and lose my job then you left for the final time.

Table 7: Summary of topics from self-stigmatizing posts. Interpersonal entities in blue, substances in green, and actions in purple.