Socio-Computational Analyses of Stigma Disclosures from People Who Use Drugs on Reddit with LLM-Based Interventions: A Thesis Proposal

Anonymous ACL submission

Abstract

This thesis proposal explores how NLP can be used to characterize, model, and mitigate the stigma associated with substance use. Drawing on social theory and lived experience, the work is organized across three aims: (1) identifying patterns of stigma expression in Reddit communities through large-scale annotation and clustering, (2) modeling internalized stigma as a temporal and narrative process using user-level affect trajectories, and (3) designing generative tools to offer stigma-responsive support. Completed studies demonstrate how theory-grounded annotation pipelines and large language models (LLMs) can surface rich typologies of stigma and self-disclosure, while proposed work extends these insights toward context-sensitive intervention. By integrating insights from public health, narrative theory, and computational social science, this research contributes new methods for understanding and responding to stigmatizing language in real-world online settings.

1 Introduction

Stigma is a powerful social process that shapes how people are perceived, treated, and included or excluded in society. For people who use drugs (PWUD), stigma is associated with reduced access to healthcare, lower treatment uptake, and worsened health outcomes (Earnshaw and Quinn, 2012; Van Brakel, 2006). Decades of research have underscored that stigma is not merely an individual phenomenon but a multidimensional one, unfolding at individual, interpersonal, and structural levels (Link and Phelan, 2001; Corrigan et al., 2005; Stangl et al., 2019). Yet despite a robust theoretical foundation, empirical work on stigma often remains siloed in clinical or survey-based research—centered primarily on treatment-engaged populations and rarely capturing the everyday realities of those who remain outside formal systems of care.

Social media platforms such as Reddit offer new opportunities to study how stigma is experienced, narrated, and negotiated in everyday life. These platforms host rich, naturally occurring disclosures of substance use (Chancellor et al., 2019; Lu et al., 2019), providing access to the perspectives of people who are often overlooked in traditional data sources. Yet despite the volume and richness of this discourse, computational work in NLP has largely focused on classification (Almeida et al., 2024) and harm detection (Hazlehurst et al., 2019) tasks. These approaches treat stigma as a static label rather than a dynamic process, leaving open questions about how stigma emerges by means of expression, its narrative structure, and the internalization process across time and social context.

This PhD Thesis Proposal takes a socio-computational approach to study how PWUD describe and negotiate experiences of stigma on Reddit. Drawing from theories in medical sociology and leveraging methods from NLP, we aim to characterize the language of stigma and explore how it evolves over time. This work contributes not only to stigma scholarship but also to the development of NLP methods for social good that center social context and user agency.

Research Aims

This thesis is organized around three aims:

- **Aim 1:** Identify and characterize substance use communities and disclosures on Reddit, with a focus on how PWUDs share experiences related to stigma on online platforms. This includes network analysis of drug-related communities, a taxonomy of personal drug-use disclosures, and a typology of stigma expressions grounded in experiential and action-oriented theories of stigma.
- **Aim 2:** Investigate how indicators of internalized stigma (e.g., shame, stereotype endorse-

ment) evolve over time through longitudinal analysis of user narratives.

- **Aim 3:** Develop and evaluate narrative-aware NLP tools that both (a) transform stigmatizing language into more empathetic alternatives and (b) generate persona-sensitive responses to self-stigmatizing disclosures. This aim builds on prior work that introduced a destigmatization framework using large language models (LLMs), and extends it to explore the feasibility of personalized support generation based on inferred user personas and support strategies.

2 Related Work

This research integrates scholarship from stigma theory, social computing, health narrativity, and language generation. While each domain contributes valuable insights, there remains a need for integrative approaches that models stigma's complexity and translate these understandings into NLP systems that are sensitive to user agency.

2.1 Stigma Theory and Substance Use

Stigma is characterized by labeling, stereotyping, separation, status loss, and discrimination within a power context (Link and Phelan, 2001). In the context of substance use, stigma manifests across multiple dimensions—enacted (experienced discrimination), anticipated (fear of future discrimination), internalized (adoption of negative stereotypes), perceived (awareness of societal attitudes), and structural (institutional policies and norms)(Stangl et al., 2019; Earnshaw, 2020). Internalized stigma is especially harmful, linked to increased shame, depression, and reduced help-seeking (Luoma et al., 2013; Corrigan and Rao, 2012). Empirical work in this domain often focuses on clinical or treatment-engaged populations, therefore there is a need for approaches that not only model these dimensions simultaneously but also take into consideration the vast narratives of those often left out of formal systems of care.

2.2 Online Health Communities and Peer Support

Pseudonymous platforms like Reddit are key spaces for individuals managing stigmatized identities (De Choudhury and De, 2014; Andalibi et al., 2017). These communities facilitate candid self-disclosure, emotional and informational support (Sharma and De Choudhury, 2018), the formation of supportive peer networks, and a sense of belonging - particularly for PWUD, who often experience exclusions within formal and informal settings (Costello et al.; Bunting et al., 2021; Bouzoubaa et al., 2024b).

Prior work has examined the types of support exchanged (e.g., validation, advice), the role of anonymity in enabling disclosure, and how peer-driven harm reduction and recovery discourse unfolds (Wombacher et al., 2020a). However, the narrative structures and stigma dynamics expressed in these communities remain underexplored, particularly from a longitudinal or psychosocial perspective.

2.3 Computational Analysis of Health Narratives and Stigma

NLP techniques are increasingly applied to health-related social media data, especially for condition detection (Gaur et al., 2018; Strapparava and Mihalcea, 2017), predicting risk (Garg et al., 2021), and classifying drug-related events (Almeida et al., 2024; Al-Garadi et al., 2021; Sarker et al., 2022; Bouzoubaa et al., 2024a). In the context of stigma, NLP has been used to identify stigmatizing language (Li et al., 2018; Chen et al., 2022; Roesler et al., 2024; Harrigian et al., 2023) and to develop annotation frameworks that draw from stigma theory (Straton et al., 2020).

However, most computational work on stigma has two key limitations: it often uses binary classification schemes (stigmatizing / not stigmatizing) and tends to isolate a single substance or form of stigma (Eschliman et al., 2024), limiting its ability to model the multidimensional and intersectional nature of stigma as conceptualized in sociological frameworks (Link and Phelan, 2001; Corrigan et al., 2011). Further, few approaches consider narrative structure (Piper et al., 2021) or how stigma is framed, contested, or internalized over time.

2.4 Language Generation for Social Good

While NLP has made significant advances in detecting harmful content such as toxicity or hate speech (Guo et al., 2023), research on *transforming* such language remains limited. Emerging research has explored text detoxification and bias mitigation by rewriting problematic content while preserving author intent (Pryzant et al., 2020), as well as using parallel corpora to neutralize stigmatizing expressions in the context of mental illness Choey (2023).

Specific to substance use, public health literature has emphasized the importance of person-first language (Kelly et al., 2010) and sympathetic framing strategies to counteract stigma and improve treatment perceptions (McGinty et al.). Although these studies are not computational in nature, they underscore the impact of language framing on attitudes and clinical outcomes.

The evolution of generative models, particularly toward context-aware and empathetic text generation (Sharma et al., 2020; Majumder et al., 2020), has opened new avenues for supportive dialogue systems. Controlled generation approaches like CTRL (Keskar et al., 2019) offer the potential for steering outputs toward affirming, non-stigmatizing language. However, tailoring such models for high-risk, user-sensitive domains like SUD remains a nascent challenge.

This dissertation builds on these efforts by designing and evaluating generative pipelines that not only reduce public stigma in online discourse but also produce personalized, empathetic responses to internalized stigma disclosures. Importantly, our work grounds these models in lived experience, stigma theory, and harm reduction principles.

3 Aim 1: Characterizing Online Substance Use Communities and Expressions of Stigma

Despite widespread recognition that stigma creates significant barriers to treatment for substance use disorders (Farhoudian et al., 2022; Rapp et al., 2006), our understanding of how stigma manifests and is expressed in naturalistic settings is limited. Traditional research methods, such as surveys and clinical interviews, often exclude individuals not engaged in formal treatment systems, omitting a large portion of lived experience (Ashford et al., 2018). To address these gaps, Aim 1 develops computational approaches to analyze Reddit as a naturalistic site of substance use discourse, with a focus on stigma expression. This aim includes three studies that span network analysis, narrative taxonomy development, and theory-informed stigma phenotyping.

3.1 Characterizing Network Structure of Substance Use Communities

Reddit's affordances make it a valuable space for observing stigmatized health discourse (De Choudhury and De, 2014; Chancellor et al., 2019). Yet few studies have examined the structure and thematic focus of drug-related subreddits at scale.

We conducted a comprehensive network analysis of 131 drug-related subreddits, capturing the interactions of 1,368 active users in order to examine both community structures and interaction patterns. Our findings showed that the continued-use subreddits we analyzed (*r/opiates*, *r/benzodiazepines*, *r/LSD*, *r/cocaine*), resemble small-world networks (Humphries and Gurney, 2008) with dense local connectivity and bridging subreddits (e.g., *r/Drugs*, *r/AskReddit*) that facilitate information flow across communities.

A key contribution of this study was the development of a revised taxonomy for classifying drug-related online communities, grounded in thematic analysis of community descriptions (Cohen's $\kappa = 0.76$). This taxonomy modernizes an earlier classification framework by Schifano et al. (2006) to reflect contemporary Reddit discourse, including new categories such as harm reduction, recovery, and experience seeking. Beyond its sociological value, the taxonomy informed sampling and stratification strategies for downstream NLP tasks involving stigma detection and discourse analysis.

3.2 Taxonomy of Personal Drug Experience Narratives

To enable NLP systems to perform fine-grained analysis and generate contextually appropriate interventions within substance use discourse, we must first systematically deconstruct how individuals articulate their lived experiences, moving beyond coarse categorizations. Prior NLP work in this space has typically treated posts as monoliths (Almeida et al., 2024), overlooking the variety of intentions and experience types embedded in user narratives. As a result, we introduced a multi-level, multi-label annotation schema capturing three dimensions: (1) connection type (inquisition vs. disclosure), (2) subject of experience (dependency, recovery, other), and (3) specific objectives (e.g., effects, methods of ingestion, safety concerns).

We annotated 500 posts and used this dataset to train and evaluate classification models, demonstrating that LLMs, specifically GPT-4 (OpenAI, 2024), prompted using an 'Instruction + Definition + Examples' (I+D+E) strategy, achieving an F1-score of 0.91 for Connection type. Applying our best-performing model to an additional 1,000 randomly selected posts, we conducted a psycholinguistic analysis using the Linguistic Inquiry and Word Count (LIWC) (Boyd et al., 2022). Non-parametric Mann-Whitney U tests (Mann and Whitney, 1947) (with Benjamini-Hochberg FDR (Benjamini and Hochberg, 1995) control across 85 LIWC categories) revealed significant linguistic differences ($p < .05$) between Inquisition vs. Disclosure posts and Recovery vs. Dependency posts. For instance, Inquisition posts contained significantly more *polite* and *authenticity*-related language, while posts inquiring about Dependency concerns used 2.54 times more *prosocial* language (e.g., "care", "help"), suggesting strategic linguistic framing to elicit support. These findings not only validate the schema but also reveal linguistic markers relevant to stigma classification and support detection.

3.3 Phenotyping Stigma Expressions

Most computational approaches to stigma rely on binary classifications (e.g., stigmatizing vs. non-stigmatizing) (Straton et al., 2020) or focus narrowly on a single dimension or substance community (Chen et al., 2022), failing to capture the multidimensional nature of stigma as theorized in sociological literature (Stangl et al., 2019; Corrigan et al., 2006). To address this limitation, we developed a theory-informed computational framework to phenotype different expressions of stigma in online discourse.

Using a large-scale Reddit corpus ($n > 1.03M$), we first applied an LLM-based filtering process to identify 56,000 posts likely to contain at least one of five stigma types: internalized, anticipated, enacted, structural, and a novel category we introduced—Stigma Perceptions and Commentary (SPC). The SPC category captures meta-discourse about stigma (e.g., posts reflecting on society's treatment of drug users or critiquing media narratives), and its inclusion was a key conceptual and methodological contribution. It enabled us to capture critical narratives that do not fit neatly into existing stigma typologies but are essential to understanding how stigma is discussed and resisted. These five categories were used as a filter to construct the dataset for clustering and downstream analysis.

To characterize how stigma is expressed within this dataset, we annotated the posts across 17 validated, theory-derived dimensions. These included indicators of stigma experience (e.g., source of stigma, identity), narrative structure (e.g., worldmaking, presence of narrative agency), and psychological impact (e.g., shame, alienation, disengagement), drawing from validated frameworks in stigma research (Luoma et al., 2013; Smith et al., 2016; Ritsher et al., 2003), clinical psychological impact (Cook, 1987; Luoma et al., 2013), and narrativity theory (Piper et al., 2021). Applying K-Means clustering to these annotated dimensions, we identified three dominant stigma phenotypes:

- Internalized: characterized by self-blame, emotional distress, and high narrative agency. These posts often describe personal struggles with drug use, shame, and feelings of worthlessness, with internalized or self-sourced stigma present in 81.7% of the cluster. These findings validate theoretical predictions about the consequences of stigma on well-being (Hatzenbuehler et al., 2013).

- Public Stigma: focused on external discrimination, often involving treatment denial, stereotyping by providers, or institutional barriers. These narratives frequently reference enacted or structural stigma.

- Righteous Indignation: consisting of posts that critique systemic injustice without including personal narratives. These were marked by analytical language, lower narrative agency, and elevated indicators of structural or anticipated stigma. This finding aligns with Corrigan's concept of the "paradox of self-stigma" (Corrigan and Rao, 2012), where some individuals respond to societal stigma not with internalization but with righteous anger directed at systemic injustice.

These three clusters, internalized, public, and righteous, resonate closely with Corrigan's progressive model of stigma, which distinguishes between self-stigma and public stigma, while also recognizing the potential for stigma resistance through indignation or advocacy (Corrigan and Rao, 2012). The emergence of these patterns from unsupervised clustering of naturalistic text data not only validates the relevance of Corrigan's model in real-world, user-generated discourse but also affirms the value of social media data in stigma research. That individuals organically express experiences and critiques of stigma in ways that reflect established theoretical constructs suggests that such narratives are meaningful psychological artifacts—suitable not only for sociological analysis but also for computational modeling. For NLP and computational social science, these findings demonstrate that large-scale text analysis can recover psychologically and socially valid constructs, supporting future efforts to build more context-sensitive models of health-related stigma and lived experience.

4 Aim 2: Modeling Internalization as a Pyschosocial and Temporal Process

While Aim 1 focused on characterizing the landscape and expressions of stigma at specific points in time, Aim 2 shifts to a longitudinal perspective; modeling internalized stigma as an evolving psychosocial process. Internalized stigma, where individuals apply negative societal beliefs about substance use to themselves, strongly impacts self-efficacy and recovery outcomes (Luoma et al., 2013; Brown et al., 2015). Although online forums can offer spaces for identity work, peer support, and resistance to stigma (Wombacher et al., 2020b; MacLean et al., 2015), little is known about how internalization emerges, persists, or wanes over time in these settings. This aim seeks to fill that gap using temporal modeling of affect and narrative features.

4.1 Modeling Internalization of Stigma Over Time

This study models internalized stigma as a latent psychological process that unfolds over time. We build on the typology developed in Aim 1 to model how internalization emerges, intensifies, or diminishes within users' longitudinal posting trajectories. This approach responds to growing calls in stigma research to treat internalization not as a static label but as a context-dependent and dynamic psychosocial state (Earnshaw and Fox, 2024; Earnshaw et al., 2022). Our dataset includes approximately 19,000 posts from 110 users who made at least one post labeled as internalized stigma ("I") in our prior stigma phenotyping study (see Section 3.3). For each user, we construct a chronological timeline of all posts made, including drug-related and recovery-oriented subreddits.

To model internalization over time, we use a hybrid approach to extract affective and contextual features from each post. Specifically, we compute post-level scores for four emotion indicators, shame, guilt, anxiety, and depression, using an adapted version of the Emotion Affective Intensity with Sentiment (EAISe) framework by Babanejad et al. (2020). This method integrates lexicon-based emotion intensity scores (Mohammad, 2018) with frequency-based weighting and expanding term coverage via WordNet-Affect (Strapparava and Valitutti, 2004).

We also annotate each post with a set of contextual and event-based features to model how internalization is shaped by social and structural factors. These include: recovery transitions (e.g., movement from *r/opiates* to *r/OpiatesRecovery*), other stigma types present in the post (e.g., enacted, anticipated, structural), and the presence or absence of MIRC themes (Bowen et al., 2023); social, human, physical, and cultural forms of recovery capital. To enable the detection of MIRC dimensions at scale, we are developing supervised classifiers trained on a MIRC-annotated subset of posts. These dimensions offer a valuable lens for understanding how expressions of internalization may co-occur with signals of resilience, support, or identity work.

We are currently exploring several modeling strategies to capture patterns of escalation, de-escalation, or persistence in internalization trajectories. These include growth curve modeling, event-based models (e.g., time to next internalized stigma post), and state-transition analysis. This work builds on prior research using NLP to investigate addiction recovery pathways from online data (Lu et al., 2019) and grounding the Transtheoretical Model of behavior change (Prochaska and Velicer, 1997) from online support narratives (MacLean et al., 2015). By integrating affective, structural, and narrative-level signals, this study aims to produce a robust computational framework for understanding the temporal dynamics of internalized stigma in naturalistic social media contexts.

4.2 Narrative Archetypes of Self-Stigmatizers

Building on the timelines constructed in Aim 2.1, this study explores whether internalized stigma manifests more as a trait—a stable, persistent identity—or as a state, fluctuating in response to situational factors. Specifically, we ask: *Do users shift into and out of internalization over time, or do some individuals remain consistently embedded in self-stigmatizing discourse, regardless of context?*

Our goal is to identify distinct internalization trajectories or archetypes by clustering users based on how these patterns evolve over time. We are currently evaluating several modeling approaches suited to temporal sequence data, including latent profile analysis, sequence clustering, hidden Markov models, and LLM-powered persona extraction pipelines (Sun et al., 2024; Tseng et al., 2024). The resulting clusters may reveal user-level differences such as: (1) Trait-like internalizers: individuals whose posts consistently reflect shame, guilt, and stereotype endorsement, even amid signals of recovery or support; (2) State-like internalizers: individuals whose internalization emerges during moments of crisis or transition but dissipates following community engagement or emotional regulation; (3) Unaffected or resilient users: individuals who rarely express internalized stigma, even when discussing experiences of marginalization or harm.

Once these archetypes are identified, we will apply narrative theory to dissect how users in each cluster construct meaning through language. Using features such as narrative agency, temporal structure, causal framing, and moral positioning (Piper et al., 2021), we will qualitatively and computationally characterize how these users frame their relationship to stigma, recovery, and community. For example, trait-like internalizers may narrate substance use as a core part of a morally degraded self, while state-like internalizers may use more situational or redemptive framing.

Expected Outcomes: The outputs of this study will include: (1) a typology of internalization personas grounded in longitudinal text data, and (2) a methodological framework for linking trajectory-based clustering with narrative structure analysis. These findings will inform the design of persona-sensitive interventions in Aim 3 and offer broader insight into the discursive lives of self-stigmatizing individuals online (see Figure 1). More broadly, this work contributes to computational social science by showing how longitudinal affect and narrative structure can reveal psychologically meaningful user-level archetypes—a step toward more empathetic and context-aware NLP systems.

5 Aim 3: Developing and Evaluating Narrative-Aware NLP Tools

The final aim of this dissertation translates the findings from Aims 1 and 2 into NLP tools that not only analyze stigma but actively work to counteract it. Existing NLP work in public health has largely focused on the detection of stigma, risk, or misinformation rather than response. Aim 3 builds on recent advances in controlled text generation, empathetic language modeling, and human-centered AI to explore how language models can support people who use drugs by reframing stigmatizing narratives and generating context-sensitive, affirming alternatives.

5.1 Destigmatizing Harmful Language and Reframe Library

While understanding stigma's manifestations is important, developing practical interventions to reduce harmful language represents a critical next step. Most computational approaches to problematic content focus on detection rather than transformation (Nobata et al., 2016; Davidson et al., 2017), creating a gap between identification and intervention.

We analyzed over 1.5 million Reddit posts from non-drug-related subreddits and identified 3,207 posts exhibiting stigmatizing language toward PWUD, primarily in the form of directed stigma (e.g., labeling, stereotyping, dehumanization). Our classification schema was grounded in Link and Phelan (2001) stigma framework and included four core components: labeling, stereotyping, separation/status loss, and discrimination. Posts were labeled using a hybrid pipeline combining LLM prompting (GPT-3.5 and GPT-4T) with manual validation and explanation layers to ensure interpretability and alignment with theory. We then developed a multi-stage rewriting pipeline for destigmatization. Across three model settings, we tested: (1) A zero-shot baseline with no contextual knowledge; (2) an Informed model, primed with definitions and examples of stigma elements derived from our theoretical schema; and (3) an Informed + Stylized model, which added controls for emotional tone, syntax, lexical diversity, and writing style using auxiliary models (e.g., RoBERTa on GoEmotions, LIWC, MTLD).

Our best-performing model, Informed + Stylized GPT-4T, was able to rewrite posts in ways that significantly reduced stigma while preserving both tone and semantic content. For example, the phrase "I have no empathy for junkies" was reframed as "I find it difficult to empathize with individuals facing substance use challenges," demonstrating both lexical transformation and softened judgment without erasing the speaker's perspective. Human evaluation ($N = 110$) showed this model ranked highest in overall quality, faithfulness, and appropriateness, outperforming both zero-shot and non-stylized alternatives. We also conducted an automatic stylistic evaluation using LIWC and paired t-tests, finding no significant difference in overall psycholinguistic features between original and

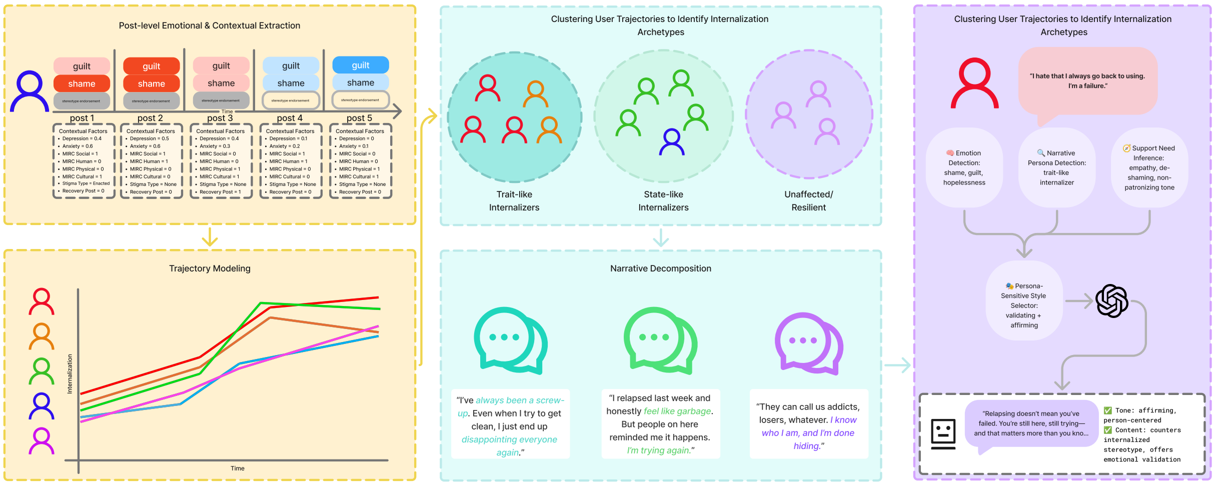
Post-level Emotional & Contextual Extraction
guilt
shame
stereotype endorsement
Time
post 1
post 2
post 3
post 4
post 5
Contextual Factors
Depression = 0.4
Anxiety = 0.6
MIRC Social = 1
MIRC Human = 0
MIRC Physical = 0
MIRC Cultural = 1
Stigma Type = Enacted
Recovery Post = 0
Contextual Factors
Depression = 0.5
Anxiety = 0.6
MIRC Social = 1
MIRC Human = 1
MIRC Physical = 0
MIRC Cultural = 0
Stigma Type = None
Recovery Post = 0
Contextual Factors
Depression = 0.4
Anxiety = 0.3
MIRC Social = 0
MIRC Human = 0
MIRC Physical = 1
MIRC Cultural = 1
Stigma Type = None
Recovery Post = 1
Contextual Factors
Depression = 0.1
Anxiety = 0.2
MIRC Social = 1
MIRC Human = 1
MIRC Physical = 0
MIRC Cultural = 0
Stigma Type = None
Recovery Post = 0
Contextual Factors
Depression = 0
Anxiety = 0.1
MIRC Social = 0
MIRC Human = 0
MIRC Physical = 1
MIRC Cultural = 1
Stigma Type = None
Recovery Post = 0
Trajectory Modeling
Internalization
Time
Clustering User Trajectories to Identify Internalization Archetypes
Trait-like Internalizers
State-like Internalizers
Unaffected/ Resilient
Narrative Decomposition
"I've always been a screw-up. Even when I try to get clean, I just end up disappointing everyone again."
"I relapsed last week and honestly feel like garbage. But people on here reminded me it happens. I'm trying again."
"They can call us addicts, losers, whatever. I know who I am, and I'm done hiding."
Clustering User Trajectories to Identify Internalization Archetypes
"I hate that I always go back to using. I'm a failure."
Emotion Detection: shame, guilt, hopelessness
Narrative Persona Detection: trait-like internalizer
Support Need Inference: empathy, de-shaming, non-patronizing tone
Persona-Sensitive Style Selector: validating + affirming
"Relapsing doesn't mean you've failed. You're still here, still trying—and that matters more than you kno...
✅ Tone: affirming, person-centered
✅ Content: counters internalized stereotype, offers emotional validation

Figure 1: Modeling of internalized stigma as a latent psychosocial process, beginning with post-level extraction of affective and contextual features to model user trajectories (far left). These insights will then be used to cluster users to identify trait-like, state-like, and resilient internalization archetypes (center). Finally, findings from these two studies will inform the development of our LLM-based supportive tool (far right).

rewritten posts—evidence that our method maintains author's voice while reducing harm. As part of this work, we also release a de-identified dataset of stigma–reframed text pairs along and a Python library designed to reduce harmful language across domains such as journalism and healthcare documentation. Importantly, this system intentionally targets public stigma directed at others rather than attempting to "correct" self-stigmatizing language used by individuals to describe their own experiences. This ethical constraint respects the agency and autonomy of people with lived experience while still addressing harmful language in broader discourse.

5.2 Persona-Conditioned Generative Support

This study builds on our prior work and proposes a persona-conditioned NLP pipeline that generates narrative-sensitive and emotionally supportive responses to self-stigmatizing disclosures. Drawing on user archetypes identified in Aim 2.2 (trait-like, state-like, resilient), we train a persona classification model using the longitudinal Reddit timelines. This model learns to associate linguistic and affective features (e.g., emotion intensity, narrative agency, self-blame framing) with specific internalization profiles. Once trained, the classifier is designed to infer a user's likely persona from their interactions.

Inspired by principles from Narrative Enhancement and Cognitive Therapy (NECT)—a validated group intervention for internalized stigma that combines psychoeducation, narrative reframing, and cognitive restructuring (Yanos et al., 2012; Roe et al., 2014) we map each persona to a support strategy profile. Rather than dramatically altering strategy types (e.g., affirmation vs. challenge), we adjust the intensity, tone, and narrative framing of shared strategies (e.g., affirmation, reflection, reappraisal) to better match the user's expressive stance. Trait-like users may benefit from more direct affirmation and belief-challenging, while state-like users may respond better to reflective validation and encouragement.

For response generation, we will fine-tune or prompt LLMs (e.g., GPT-4o (OpenAI et al., 2024), Llama (Touvron et al., 2023)) to produce responses conditioned on: (1) the original post, (2) inferred persona, and (3) target support strategy. Building on our methods in Aim 3.1, we incorporate stylistic conditioning techniques to match emotional tone, sentence complexity, and narrative framing. We will also explore contrastive learning or reward-based alignment (e.g., RLHF) to guide persona-response alignment.

Evaluation: We frame this study as a conceptual evaluation of how well LLMs can be guided to produce supportive responses that express linguistic features associated with reduced self-stigma, perceived social support, and empowerment. We do not make claims about long-term behavioral or psychological outcomes; rather, we assess the degree to which generated responses reflect key discourse-level properties linked to effective stigma-reducing interventions (Rüsch et al., 2014; Yanos et al., 2015). Human evaluation will be conducted with a small panel of trained annotators, including harm reduction experts and clinical psychologists. Annotators will assess generated responses in terms of empathy, narrative appropriateness (Hu et al., 2024; Lei et al., 2025), and the extent to which the system respectfully challenges internalized negative beliefs without erasing the user's agency or voice. These judgments will provide a grounded sense of whether the system expresses discourse-level qualities aligned with stigma reduction and supportive communication. We will also report basic automatic metrics such as fluency and perplexity while acknowledging the limitations of current evaluation methods for capturing complex social attributes like empathy and empowerment (Schmidtova et al., 2024). We will also conduct an A/B comparison between persona-conditioned and generic empathetic responses and optionally include an ethics audit to examine model behavior for risks such as unintended condescension, narrative mismatch, or reinforcing stigma.

6 Conclusion

This thesis proposal contributes to the growing field of socially informed NLP and NLP for social good by developing methods that characterize, model, and intervene in the language of stigma surrounding substance use. Across three aims, it brings together theory-grounded annotation, temporal modeling of psychosocial states, and generative NLP systems to support PWUD. By integrating lived experience and narrative structure into each phase, this work not only advances our understanding of how stigma is expressed and internalized in online settings but also explores how language models can be used to promote more humane and empathetic responses to those expressions.

7 Ethics Statement

7.1 Data Sources and Consent

This research uses publicly available Reddit posts related to substance use. All data collection follows ethical guidelines for working with social media data, including respecting platform terms of service and user anonymity (Moreno et al., 2013). Posts are analyzed without user handles or identifiable metadata. Because the data are public and observational, this work did not require institutional IRB approval, though it adheres to institutional guidelines for secondary data analysis.

7.2 Risk Mitigation and Harm Reduction

Given the sensitive nature of substance use discourse and stigma, special care is taken to avoid reproducing harm. Analyses of stigmatizing content are contextualized within sociological and public health frameworks to avoid deficit framing. For generation tasks, model outputs are reviewed to avoid condescension, overcorrection, or erasure of user voice. Future iterations will include participatory audits with harm reduction experts and individuals with lived experience.

7.3 Researcher Positionality and Motivation

This work is grounded in a public health background, shaped by prior research experience with large-scale substance use studies like the NIH HEAL Initiative's Healing Communities Study[1]. I worked with rich clinical and behavioral data to examine risk trajectories for opioid relapse, often accompanied by unstructured text such as clinical notes or patient narratives. What stood out was how rarely these narratives were treated as meaningful data sources in their own right—particularly with regard to how individuals described stigma, identity, and care experiences.

This research emerged from a desire to center those narratives—to move beyond predictive modeling toward interpretive, context-aware approaches that treat language as both data and a form of expression. I see this work not only as a technical contribution but as a way to reshape how we design computational systems that engage with marginalized communities.

8 Limitations

This work has several limitations. First, while Reddit offers rich, candid narratives from people who use drugs, its user base is not demographically representative, which may limit the generalizability of findings to broader populations. Second, the use of LLMs, while powerful for generation and annotation, introduces challenges in interpretability, particularly when modeling complex psychosocial phenomena like stigma.

References

Mohammed Ali Al-Garadi, Yuan-Chi Yang, Haitao Cai, Yucheng Ruan, Karen O'Connor, Gonzalez-Hernandez Graciela, Jeanmarie Perrone, and Abeed

[1] https://hcs.rti.org/

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.

Alexandra Almeida, Thomas Patton, Mike Conway, Amarnath Gupta, Steffanie A. Strathdee, and Annick Bórquez. 2024. The Use of Natural Language Processing Methods in Reddit to Investigate Opioid Use: Scoping Review. *JMIR Infodemiology*, 4(1):e51156. Company: JMIR Infodemiology Distributor: JMIR Infodemiology Institution: JMIR Infodemiology Label: JMIR Infodemiology Publisher: JMIR Publications Inc., Toronto, Canada.

Nazanin Andalibi, Pinar Ozturk, and Andrea Forte. 2017. Sensitive Self-disclosures, Responses, and Social Support on Instagram: The Case of #Depression. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 1485–1500, New York, NY, USA. Association for Computing Machinery.

Robert D. Ashford, Austin M. Brown, and Brenda Curtis. 2018. Systemic barriers in substance use disorder treatment: A prospective qualitative study of professionals in the field. *Drug and Alcohol Dependence*, 189:62–69.

Nastaran Babanejad, Heidar Davoudi, Aijun An, and Manos Papagelis. 2020. Affective and contextual embedding for sarcasm detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, page 225–243, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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.

Layla Bouzoubaa, Elham Aghakhani, Max Song, Quang Trinh, and Shadi Rezapour. 2024a. Decoding the narratives: Analyzing personal drug experiences shared on Reddit. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6131–6148, Bangkok, Thailand. Association for Computational Linguistics.

Layla Bouzoubaa, Jordyn Young, and Rezvaneh Rezapour. 2024b. 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*, ASONAM '23, pages 558–565, New York, NY, USA. Association for Computing Machinery.

Elizabeth Bowen, Amanda Irish, Gregory Wilding, Chelsea LaBarre, Nicholas Capozziello, Thomas Nochajski, Robert Granfield, and Laura A. Kaskutas. 2023. Development and psychometric properties of the multidimensional inventory of recovery capital (mirc). *Drug and Alcohol Dependence*, 247:109875.

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.

Seth A. Brown, Kirstin Kramer, Brittany Lewno, Luci Dumas, Gina Sacchetti, and Elisa Powell. 2015. Correlates of Self-Stigma among Individuals with Substance Use Problems. *International Journal of Mental Health and Addiction*, 13(6):687–698.

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.

Stevie Chancellor, George Nitzburg, Andrea Hu, Francisco Zampieri, and Munmun De Choudhury. 2019. Discovering Alternative Treatments for Opioid Use Recovery Using Social Media. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–15, New York, NY, USA. Association for Computing Machinery.

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.

Mason Choey. 2023. From stigma to support: A parallel monolingual corpus and NLP approach for neutralizing mental illness bias. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 249–254, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Duane R. Cook. 1987. Measuring shame: The internalized shame scale. *Alcoholism Treatment Quarterly*, 4(2):197–215.

Patrick W. Corrigan, Amy Kerr, and Lissa Knudsen. 2005. The stigma of mental illness: Explanatory models and methods for change. *Applied and Preventive Psychology*, 11(3):179–190.

Patrick W. Corrigan, Jennifer Rafacz, and Nicolas Rüsch. 2011. Examining a progressive model of self-stigma and its impact on people with serious mental illness. *Psychiatry Research*, 189(3):339–343.

Patrick W. Corrigan and Deepa Rao. 2012. On the self-stigma of mental illness: Stages, disclosure, and strategies for change. *Canadian journal of psychiatry. Revue canadienne de psychiatrie*, 57(8):464–469.

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):239–246. Place: US Publisher: American Psychological Association.

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

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *International Conference on Web and Social Media*.

Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 71–80.

Valerie A. Earnshaw. 2020. Stigma and substance use disorders: A clinical, research, and advocacy agenda. *The American psychologist*, 75(9):1300–1311.

Valerie A Earnshaw and Annie B Fox. 2024. Advancing substance use disorder stigma research: It's about time. *Stigma and Health*.

Valerie A. Earnshaw and Diane M. Quinn. 2012. The Impact of Stigma in Healthcare on People Living with Chronic Illnesses. *Journal of Health Psychology*, 17(2):157–168. Publisher: SAGE Publications Ltd.

Valerie A Earnshaw, Ryan J Watson, Lisa A Eaton, Natalie M Brousseau, Jean-Philippe Laurenceau, and Annie B Fox. 2022. Integrating time into stigma and health research. *Nature Reviews Psychology*, 1(4):236–247.

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.

Ali Farhoudian, Emran Razaghi, Zahra Hooshyari, Alireza Noroozi, Azam Pilevari, Azarakhsh Mokri, Mohammad Reza Mohammadi, and Mohsen Malekinejad. 2022. Barriers and Facilitators to Substance Use Disorder Treatment: An Overview of Systematic Reviews. *Substance Abuse: Research and Treatment*, 16:11782218221118462.

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.

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

Keith Harrigian, Ayah Zirikly, Brant Chee, Alya Ahmad, Anne Links, Somnath Saha, Mary Catherine Beach, and Mark Dredze. 2023. Characterization of stigmatizing language in medical records. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 312–329, Toronto, Canada. Association for Computational Linguistics.

Mark L. Hatzenbuehler, Jo C. Phelan, and Bruce G. Link. 2013. Stigma as a Fundamental Cause of Population Health Inequalities. *American Journal of Public Health*, 103(5):813–821.

Brian Hazlehurst, Carla A Green, Nancy A Perrin, John Brandes, David S Carrell, Andrew Baer, Angela DeVeaugh-Geiss, and Paul M Coplan. 2019. Using natural language processing of clinical text to enhance identification of opioid-related overdoses in electronic health records data. *Pharmacoepidemiology and drug safety*, 28(8):1143–1151.

Yuxuan Hu, Minghuan Tan, Chenwei Zhang, Zixuan Li, Xiaodan Liang, Min Yang, Chengming Li, and Xiping Hu. 2024. Aptness: Incorporating appraisal theory and emotion support strategies for empathetic response generation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 900–909, New York, NY, USA. Association for Computing Machinery.

Mark D Humphries and Kevin Gurney. 2008. Network 'small-world-ness': a quantitative method for determining canonical network equivalence. *PloS one*, 3(4):e0002051.

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.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *Preprint*, arXiv:1909.05858.

Xixi Lei, Changqun Li, Liang He, and Xin Lin. 2025. An interactive evaluation framework for empathetic response generation. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Ang Li, Dongdong Jiao, and Tingshao Zhu. 2018. Detecting depression stigma on social media: A linguistic analysis. *Journal of affective disorders*, 232:358–362.

Bruce G. Link and Jo C. Phelan. 2001. Conceptualizing stigma. *Annual Review of Sociology*, 27(1):363–385.

John Lu, Sumati Sridhar, Ritika Pandey, Mohammad Al Hasan, and Georege Mohler. 2019. Investigate Transitions into Drug Addiction through Text Mining of Reddit Data. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pages 2367–2375, New York, NY, USA. Association for Computing Machinery.

Jason B. Luoma, Richard H. Nobles, Chad E. Drake, Steven C. Hayes, Alyssa O'Hair, Lindsay Fletcher, and Barbara S. Kohlenberg. 2013. Self-Stigma in Substance Abuse: Development of a New Measure. *Journal of Psychopathology and Behavioral Assessment*, 35(2):223–234.

Diana MacLean, Sonal Gupta, Anna Lembke, Christopher Manning, and Jeffrey Heer. 2015. Forum77: An analysis of an online health forum dedicated to addiction recovery. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '15, pages 1511–1526, New York, NY, USA. Association for Computing Machinery.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. MIME: MIMicking emotions for empathetic response generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8968–8979, Online. Association for Computational Linguistics.

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.

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

Saif M. Mohammad. 2018. Word affect intensities. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*, Miyazaki, Japan.

Megan A Moreno, Natalie Goniu, Peter S Moreno, and Douglas Diekema. 2013. Ethics of social media research: Common concerns and practical considerations. *Cyberpsychology, behavior, and social networking*, 16(9):708–713.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 145–153, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI, : Aaron Hurst, Adam Lerer, Adam P. Goucher, and et al. Adam Perelman. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Andrew Piper, Richard Jean So, and David Bamman. 2021. Narrative Theory for Computational Narrative Understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 298–311, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

James O. Prochaska and Wayne F. Velicer. 1997. The transtheoretical model of health behavior change. *American Journal of Health Promotion*, 12(1):38–48.

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.

Richard C. Rapp, Jiangmin Xu, Carey A. Carr, D. Tim Lane, Jichuan Wang, and Robert Carlson. 2006. Treatment barriers identified by substance abusers assessed at a centralized intake unit. *Journal of substance abuse treatment*, 30(3):227–235.

Jennifer Boyd Ritsher, Poorni G Otilingam, and Monica Grajales. 2003. Internalized stigma of mental illness: psychometric properties of a new measure. *Psychiatry research*, 121(1):31–49.

David Roe, Ilanit Hasson-Ohayon, Michal Mashiach-Eizenberg, Oren Derhy, Paul H Lysaker, and Philip T Yanos. 2014. Narrative enhancement and cognitive therapy (nect) effectiveness: A quasi-experimental study. *Journal of clinical psychology*, 70(4):303–312.

David Roesler, Shana Johnny, Mike Conway, and Annie T. Chen. 2024. A theory-informed deep learning approach to extracting and characterizing substance use-related stigma in social media. *BMC Digital Health*, 2(1):60.

Nicolas Rüsch, Elvira Abbruzzese, Eva Hagedorn, Daniel Hartenhauer, Ilias Kaufmann, Jan Curschellas, Stephanie Ventling, Gianfranco Zuaboni, René Bridler, Manfred Olschewski, and 1 others. 2014. Efficacy of coming out proud to reduce stigma's impact among people with mental illness: pilot randomised controlled trial. *The British journal of psychiatry*, 204(5):391–397.

Abeed Sarker, Mohammed Ali Al-Garadi, Yao Ge, Nisha Nataraj, Christopher M. Jones, and Steven A. Sumner. 2022. Signals of increasing co-use of stimulants and opioids from online drug forum data. *Harm Reduction Journal*, 19(1):1–12. Number: 1 Publisher: BioMed Central.

Fabrizio Schifano, Paolo Deluca, Alex Baldacchino, Teuvo Peltoniemi, Norbert Scherbaum, Marta Torrens, Magi Farrě, Irene Flores, Mariangela Rossi, Dorte Eastwood, Claude Guionnet, Salman Rawaf, Lisa Agosti, Lucia Di Furia, Raffaella Brigada, Aino Majava, Holger Siemann, Mauro Leoni, Antonella Tomasin, and 2 others. 2006. Drugs on the web; the Psychonaut 2002 EU project. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 30(4):640–646.

Patricia Schmidtova, Saad Mahamood, Simone Balloccu, Ondrej Dusek, Albert Gatt, Dimitra Gkatzia, David M. Howcroft, Ondrej Platek, and Adarsa Sivaprasad. 2024. Automatic metrics in natural language generation: A survey of current evaluation practices. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 557–583, Tokyo, Japan. Association for Computational Linguistics.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.

Eva Sharma and Munmun De Choudhury. 2018. Mental Health Support and its Relationship to Linguistic Accommodation in Online Communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 1–13, New York, NY, USA. Association for Computing Machinery.

Laramie R Smith, Valerie A Earnshaw, Michael M Copenhaver, and Chinazo O Cunningham. 2016. Substance use stigma: Reliability and validity of a theory-based scale for substance-using populations. *Drug and alcohol dependence*, 162:34–43.

Anne L. Stangl, Valerie A. Earnshaw, Carmen H. Logie, Wim van Brakel, Leickness C. Simbayi, Iman Barré, and John F. Dovidio. 2019. The Health Stigma and Discrimination Framework: a global, crosscutting framework to inform research, intervention development, and policy on health-related stigmas. *BMC Medicine*, 17(1):31.

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.

Carlo Strapparava and Alessandro Valitutti. 2004. WordNet-Affect: An Affective Extension of WordNet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, pages 1083–1086. European Language Resources Association (ELRA).

Nadiya Straton, Hyeju Jang, and Raymond Ng. 2020. Stigma annotation scheme and stigmatized language detection in health-care discussions on social media. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1178–1190, Marseille, France. European Language Resources Association.

Chenkai Sun, Ke Yang, Revanth Gangi Reddy, Yi R. Fung, Hou Pong Chan, Kevin Small, ChengXiang Zhai, and Heng Ji. 2024. Persona-db: Efficient large language model personalization for response prediction with collaborative data refinement. *Preprint*, arXiv:2402.11060.

Hugo Touvron, Thibaut Lavril, and Gautier Izacard et al. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in LLMs: A survey of role-playing and personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631, Miami, Florida, USA. Association for Computational Linguistics.

Wim H. Van Brakel. 2006. Measuring health-related stigma–a literature review. *Psychology, Health & Medicine*, 11(3):307–334.

Kevin Wombacher, Sarah E. Sheff, and Natalie Itrich. 2020a. 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.

Kevin Wombacher, Sarah E. Sheff, and Natalie Itrich. 2020b. Social support for active substance users: A content analysis of r/drugs. *Health Communication*, 35(6):756–765.

Philip T Yanos, Alicia Lucksted, Amy L Drapalski, David Roe, and Paul Lysaker. 2015. Interventions targeting mental health self-stigma: A review and comparison. *Psychiatric rehabilitation journal*, 38(2):171.

Philip T Yanos, David Roe, Michelle L West, Stephen M Smith, and Paul H Lysaker. 2012. Group-based treatment for internalized stigma among persons with severe mental illness: findings from a randomized controlled trial. *Psychological services*, 9(3):248.