Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

Lived Experience Matters: Automatic Detection of Stigma toward People Who Use Substances on Social Media

Salvatore Giorgi[1,2], Douglas Bellew[1], Daniel Roy Sadek Habib[1], João Sedoc[3], Chase Smitterberg[1], Amanda Devoto[1], McKenzie Himelein-Wachowiak[1], Brenda Curtis[1]

[1] National Institute on Drug Abuse
[2] University of Pennsylvania
[3] New York University
sal.giorgi@nih.gov, brenda.curtis@nih.gov

Abstract

Stigma toward people who use substances (PWUS) is a leading barrier to seeking treatment. Further, those in treatment are more likely to drop out if they experience higher levels of stigmatization. While related concepts of hate speech and toxicity, including those targeted toward vulnerable populations, have been the focus of automatic content moderation research, stigma and, in particular, people who use substances have not. This paper explores stigma toward PWUS using a data set of roughly 5,000 public Reddit posts. We performed a crowd-sourced annotation task where workers are asked to annotate each post for the presence of stigma toward PWUS and answer a series of questions related to their experiences with substance use. Results show that workers who use substances or know someone with a substance use disorder are more likely to rate a post as stigmatizing. Building on this, we use a supervised machine learning framework that centers workers with lived substance use experience to label each Reddit post as stigmatizing. Modeling person-level demographics in addition to comment-level language results in a classification accuracy (as measured by AUC) of 0.69 – a 17% increase over modeling language alone. Finally, we explore the linguist cues which distinguish stigmatizing content: PWUS substances and those who don't agree that language around othering ("people", "they") and terms like "addict" are stigmatizing, while PWUS (as opposed to those who do not) find discussions around specific substances more stigmatizing. Our findings offer insights into the nature of perceived stigma in substance use. Additionally, these results further establish the subjective nature of such machine learning tasks, highlighting the need for understanding their social contexts.

Introduction

In the U.S. in 2021, 61.2 million people aged 12 or older (22% of the population) used substances, and 46.3 million (17% of the population) met the criteria for having a substance use disorder (SUD) (Abuse and Administration 2021). Despite the prevalence of SUDs and substance use, 94% of people with a SUD did not receive treatment. There are significant barriers to seeking treatment, including stigma (negative biases including stereotypes, prejudice, and discrimination ; Corrigan and Watson 2002). Studies have shown that, of those people who did perceive a need for treatment, 22.7% report stigma as a reason for lack of seeking out treatment (Ashford, Brown, and Curtis 2019). Stigma has measurable consequences on the health and well-being of people who use substances (PWUS): it contributes to diminished help-seeking (Clement et al. 2015), medication non-adherence (Sirey et al. 2001), healthcare avoidance (Byrne 2008), worse healthcare (Van Boekel et al. 2013), poor health outcomes (Byrne 2008; Stangl et al. 2019), and lower quality of life (Cheng et al. 2019). Indeed, stigma, in general, is a central driving force for population mortality (Hatzenbuehler, Phelan, and Link 2013).

While extensive research has been conducted on the automatic detection of related concepts of hate speech and toxic language on social media (Fortuna and Nunes 2018), stigma and, in particular, stigma towards PWUS has received relatively little attention. This is despite the fact that roughly half of the people in treatment for SUDs have reported that their online communities contain triggering content (Ashford, Lynch, and Curtis 2018). The standard pipeline for the automatic detection of hate speech on social media is to collect a corpus of posts (from Twitter or Reddit, for example) and label each post as to whether or not it contains hate speech through a crowd-sourcing annotation task. Typically produced through a majority vote across the annotations, these labels are then used to train a machine learning classifier to detect hate speech on unseen data automatically. Recently, several issues have been identified with this pipeline where the annotation process and majority voting introduce substantial biases into the final machine learning model through, for example, annotator demographics (Díaz et al. 2018), annotator beliefs (Sap et al. 2022), insensitivity to dialects of minority populations (Sap et al. 2019) Thus, increasing focus is being given to understanding *who* is annotating data, what are the annotators' beliefs, moral values, and lived experiences, and how can machine learning methods incorporate dissenting opinions and disagreement (Davani, Díaz, and Prabhakaran 2022; Prabhakaran, Davani, and Diaz 2021; Rottger et al. 2022). See Uma et al. (2021) for a survey and in-depth discussion on disagreement in learning tasks.

In this paper, we attempt to automatically identify stigmatizing content on social media by centering people who are the subject of the stigma – those who have lived experience

with substance use. This is done in an attempt to understand both manifestations of stigma and *who* perceives it through three research questions:

- **RQ1** Can stigma towards people who use substances (PWUS) be automatically identified?
- **RQ2** Does lived experience with substance use inform how stigma is perceived?
- **RQ3** Are there linguistic differences between stigma perceived by people with lived experience with substance use and those without?

To do this, we collect a sample of 5,000 Reddit comments that contain mentions of substances or substance use and run a crowd-sourcing task where we pay Amazon Mechanical Turk (Mturk) workers to label the posts as having stigma towards PWUS. We also ask the workers a series of demographic and substance use-related questions. Using this demographic information, we assign stigma labels to each Reddit comment through a learning framework that allows one to center various demographic distributions, also known as jury learning (Gordon et al. 2022). We end by examining the linguistic cues associated with stigma across different populations.

Contributions Our key contributions include: (1) the public release of a data set of stigma-annotated Reddit comments along with demographic variables of the annotators[1]; (2) we show that annotators with lived experience with substance use are more likely to label a social media post as stigmatizing through the evaluation of a machine learning classifier which centers groups of annotators with common attributes; and (3) we identify linguistic markers associated with stigmatizing social media posts as highlighted by annotators with lived experience with substance use.

Related Work

Definition of Stigma Toward People Who Use Substances Stigma can be thought of as a collection of negative biases against certain groups of people, which often incorporate three components: stereotypes, prejudice, and discrimination (Corrigan and Watson 2002). All three components can manifest through interpersonal interactions or intrapersonally, known as self-stigma. Following Link and Phelan (2001), stigma consists of the "identification of differentness, the construction of stereotypes, the separation of labeled persons into distinct categories, and the full execution of disapproval, rejection, exclusion, and discrimination" by people with access to "social, economic, and political power".

While many group experience stigma, this paper is focused on stigma experienced by people who use substances or people with a SUD. In a population of people in treatment for SUD, approximately 74% had social media and 47% reported their online communities to contain triggering content (Ashford, Lynch, and Curtis 2018). Similarly, studies have shown up to 60% of people felt they were treated unfairly due to having a SUD, and 39.5% reported at least three types of stigmatizing experiences in their daily lives (Luoma et al. 2007).

Despite widespread stigma in our society, there has been little agreement as to what constitutes stigmatizing language. Researchers have called for a standardized collection of terms or phrases (Kelly 2004) which can be utilized to better serve as an assessment tool for the understanding and sensitivity regarding mental health, SUD, and its stigma. Additionally, the use of medically appropriate language by physicians and the general population can combat stigmatizing attitudes, offering respect for people with SUD (Kelly, Wakeman, and Saitz 2015).

Stigma and Hate Speech on Social Media The first tools used to counter stigmatizing and hateful posts online put the onus on social media users themselves to label posts as inappropriate (Kayes et al. 2015). Research efforts similarly utilized manually annotated data sets to label hateful and stigmatizing social media content (Golbeck et al. 2017; McNeil, Brna, and Gordon 2012; Founta et al. 2018; Davidson et al. 2017). Studying hate speech in online text, particularly social media such as Facebook or Twitter (MacAvaney et al. 2019), has proved informative along the lines of gender (Waseem 2016; Basile et al. 2019), religion (Albadi, Kurdi, and Mishra 2018), race (De Gibert et al. 2018; Waseem and Hovy 2016), and immigration status (Basile et al. 2019; Ross et al. 2017).

Automatic Detection of Stigma Along with physical health conditions such as COVID-19 (Liu et al. 2022), automatic stigma detection has recently been implemented in the context of mental health conditions such as depression, schizophrenia, and suicide (Li et al. 2018, 2020; Jilka et al. 2022; Li, Jiao, and Zhu 2018; Oscar et al. 2017). Automatic detection has also been applied to labeling social media content related to substance use (Roy et al. 2017; Zhang et al. 2018). However, stigma toward PWUS on social media is only beginning to be explored.

Perhaps closest to the present work, Chen, Johnny, and Conway (2022) explore experiences of stigma, posted to the Reddit platform, by people who use substances. This work focuses on three types of stigma (anticipated, internalized, and enacted) and three substances (alcohol, cannabis, and opioids). While this paper also uses Reddit data and natural language processing techniques to understand stigma, it focuses on experiences of stigma as opposed to identifying stigmatizing content, which is the focus of the current study.

Data

Annotation Data

We begin with 1.66 billion Reddit comments from 2019 collected from pushshift.io (Baumgartner et al. 2020). We then identify comments which contain at least one substance use keyword (see below for keyword selection and disambiguation process) for a total of 9.3 million comments. From this, we select 5,000 random comments for our annotation task. In Table 1, we break down the substance keyword distribution of these 5,000 comments.

[1]https://github.com/TTRUCurtis/lived-experience

Keyword	N (%)	Keyword	N (%)	Keyword	N (%)	Keyword	N (%)	Keyword	N (%)
acid	379 (7.6)	dab	77 (1.5)	lsd	137 (2.7)	opiate	127 (2.5)	shrooms	73 (1.5)
adderall	78 (1.6)	drug	2396 (47.9)	marijuana	174 (3.5)	opioid	109 (2.2)	valium	19 (0.4)
addy	12 (0.2)	fentanyl	35 (0.7)	mdma	76 (1.5)	oxy	24 (0.5)	weed	762 (15.2)
cocaine	128 (2.6)	heroin	173 (3.5)	meth	216 (4.3)	oxycodone	5 (0.1)	xanax	61 (1.2)
codeine	18 (0.4)	kratom	111 (2.2)	molly	56 (1.1)	percocet	4 (0.1)	xans	16 (0.3)
coke	242 (4.8)	kush	28 (0.6)	norco	3 (0.1)	purp	7 (0.1)	xtc	6 (0.1)

Table 1: Percentage of comments in the combined training and test data containing each substance keyword.

Substance Keywords We identify comments related to substance use by identifying posts containing substance related keywords. These keywords were chosen to identify posts about specific substances (e.g., LSD or meth), a breadth of substances (e.g., we do not focus solely on opioids), general substances (e.g., drug*), and substance *use* (e.g., smoke). The Drug Enforcement Administration slang word list was used as a starting point for choosing substance keywords (Administration et al. 2018) and all keywords were agreed upon by an interdisciplinary team of substance use researchers. As a quality control check, we manually checked a random set of 1,000 posts in order to identify any obvious inconsistencies with our keyword data. This was an iterative process where these manual check were discussed as a group and they keywords were further refined. Through this process, we identified several simple heuristics designed to reduce false positives (i.e., comments that do not refer to substances) and, thus, removed comments containing the following phrases: hillary, clinton, obama, bernie, bern, sanders, trump, gab, weed out, crack jokes, crack me up, *white pill, black pill, red pill, blue pill, *whitepill*, *blackpill*, *redpill, *bluepill* and crazy pill. This resulted in 8,798,160 comments and is referred to as the Substance Keyword data set below.

Substance Keyword Disambiguation We note that multiple keywords used to identify the Substance Keyword data set have multiple senses, many of which are not related to substances. For example, "I smoked pot" and "I used a pot to cook." Thus, we attempt to refine our keyword list by removing keywords less likely to be referring to substances. To do this, two annotators were asked to rate 1,000 random comments (from the Substance Keyword data set above) for the following: "Is this post about substances? Posts may reference substances by name, slang, or you may be able to determine by context." Both annotators are substance use researchers. The two raters agreed on 93.2% of posts with a Cohen's Kappa score of 0.85. A total of 61.1% of the 1,000 posts referred to substances (where both annotators agreed). Keywords were retained if they were used to discuss substances in more than 50% of their occurrences or if they did not occur in the 1,000 random posts (using the assumption that these were rare words that would not dramatically increase false positives). This included: barbs, blunt, crack, ecstasy, joint, pot, and tabs. The keywords "acid" was found to refer to substances in only 41% of posts containing that keyword, yet, after internal discussions, it was decided that this keyword should be retained. The final list of keywords is shown in Table 1.

Model Evaluation Data

Train and Test Split The jury learning framework is an annotator-level model, which predicts each worker's annotations given comment text, past annotations, and group-level information. As such, we evaluate the model on unseen (or held out) Reddit *comments*, as opposed to unseen workers or annotations. As described below in the Annotation Task section, the final annotated data set consists of 6,147 annotations of 3,802 comments from 400 workers. We create a random train/test split by taking 80% and 20% of the comments for training and testing, respectively. This results in 4,761 annotations across 3,042 comments in the training data and 1,386 annotations across 760 comments in the test data. This data set is used for **RQ1**.

Evaluating Stigma in the Wild In order to identify stigmatizing language across Reddit (**RQ2** and **RQ3**), we apply the trained jury learning model to a large, random sample of unseen comments. As such, we collect 10,000 random comments from the Substance Keyword data set which do not appear in annotated data (i.e., do not appear in the training or testing data).

Annotation Task

We begin by asking consenting Amazon Mechanical Turk (MTurk) workers a series of demographic questions (age,

	Full Sample	Final Sample
Age	38.4 (10.5)	38.8 (10.8)
Gender		
Female	241 (42.7%)	180 (45.0%)
Male	371 (56.5%)	215 (53.8%)
Transgender, etc.	5 (0.80%)	5 (1.20%)
Race / Ethnicity		
African American	80 (14.2%)	61 (10.8%)
Asian	23 (4.1%)	20 (3.5%)
White	447 (79.1%)	308 (77.0%)
Substance Use		
Know someone	382 (67.6%)	259 (64.8%)
Use substances	369 (65.3%)	212 (37.5%)
Days of SU	6.6 (10.4)	5.0 (9.7)

Table 2: Demographic distribution of MTurk workers, Mean (SD) or # (%). Full Sample $N = 565$, Final $N = 400$.

gender identity, and race/ethnicity), how many times they have used substances for non-medical reasons within the past 30 days, and if they know anyone who is in treatment for substance use disorder. Note that due to the potentially sensitive nature of these questions, we did not force a response to the survey, which has implications for the final sample (see below for final data filtering). We then give the workers a short training on the task, specifically what types of posts reference drugs (e.g., posts about drug stores or pharmacies should not be considered) and how to define stigma. Workers are then given a short three-question quiz, where they are shown the correct answer upon completion of the quiz (no workers are removed for incorrect quiz answers). See the Appendix for the full survey question text, quiz questions, and attention check.

After completing the demographic survey and training materials, workers are then shown a series of 20 random Reddit comments. For each of the 20 comments, the workers are asked the question *Q-Sub*: "Are the drug-related words in this post being used to talk about drugs? (Yes/No)" If the worker responded *Yes*, then the second question *Q-Stigma* is asked: "Does this post contain stigmatizing language? (Yes/No)". This second question was skipped if the worker responded *No* to *Q-Sub*. For both questions, there is a 5-second delay before the submit button is displayed to force the worker to read the Reddit comment thoroughly. After 10 Reddit comments, an attention check question is asked (see Appendix for details on the attention check), which is immediately followed by the final 10 Reddit comments. The maximum number of annotations per comment was set to 3.

Workers are paid $2.50 for completing the demographic questions and annotating 20 Reddit posts (based on a $15 hourly rate). In order to both view and work on this task, workers were required to be located in the U.S., have an approval rating of 80% or higher, and have at least 100 approved HITs (Human Intelligence Tasks).

Annotation Results

Due to the randomization process and workers not completing the full task, not all of the 5,000 comments were seen by workers, nor were all comments rated three times (our desired number of annotations per comment). As such, at the end of the annotation process, a total of 4,991 comments were rated at least once by one of the 704 workers who attempted this task. We then removed workers who: (1) failed the attention check, (2) did not answer all questions in the demographic survey, and (3) did not complete the full series of 20 annotations in a single HIT (i.e., quit the task early). Note that if the worker answered *No* to *Q-Sub* then they were not asked to rate the comment for stigma. Thus, we further refined the data set to only those comments which were annotated for stigma at most three times. This produced a data set of 4,600 comments rated at least once for stigma by 565 workers for a total of 9,392 annotations. The demographic distribution of this worker sample is seen in the Full Sample column of Table 2.

Next, we examined the distribution of positive stigma labels (i.e., stigma is present in the comment) across the workers. In Figure 1, we see the percentage of labeled posts

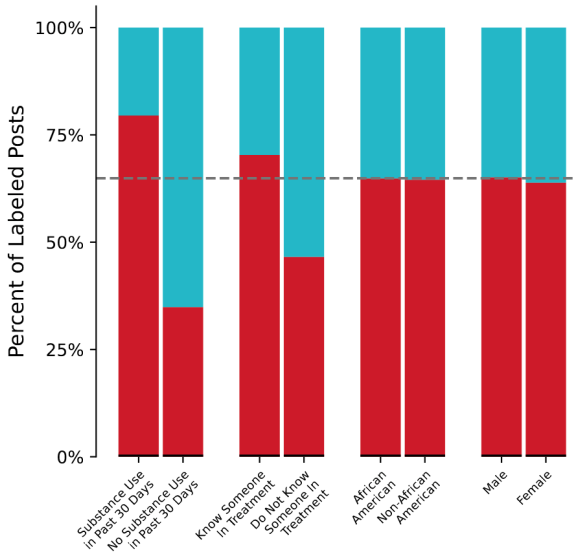

Figure 1: Percentage of stigma labels (red) and non-stigma labels (blue) for each binary demographic group across the Full Sample. Grey dotted line is the percentage of positive stigma labels across the entire sample (65%).

across each of the different demographic groups: those who use substances, those who know someone in treatment for a SUD, gender, and race/ethnicity. We see significant differences (via two-sided t-test) in the number of positive stigma labels across those who use/do not use substances ($t = 48.4$, $p < 0.01$) and those who know/don't know someone in treatment ($t = 16.4, p < 0.01$). Gender ($t = 0.62, p > 0.05$) and race/ethnicity ($t = -0.58, p > 0.05$) differences are not statistically different. In Figure 2, we plot a distribution of the percentage of positive stigma labels across each worker's total annotations (the number of positive stigma labels divided by the worker's total number of annotations). We plot

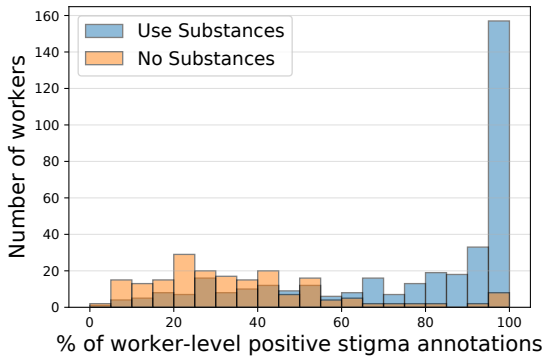

Figure 2: Distribution of the worker-level percentage of positive stigma annotations across the Full Sample (the number of positive stigma annotations divided by the worker's total number of annotations). PWUS are shown in blue and those who have not used substances are shown in orange.

	Krippendorff's α
All	.12
Gender	
Female	.14
Not Female	.13
Race / Ethnicity	
African American	.16
Not African American	.13
Knowing Someone in Treatment	
Know Someone	.12
Do Not Know Someone	.18
Substance Use in the Past 30 Days	
Substance Use	.09
No Substance Use	.31

Table 3: Krippendorff's α for each subgroup.

this distribution for both people who have used substances at least once in the past 30 days (blue) and those who did not (orange). As seen in this figure, there is a large spike at the tail end where workers rated over 95% of their annotations as stigmatizing. Notably, the majority of people in this bin use substances. While this behavior may point to bad data (random or unreliable annotations), we note that the workers in this bin passed the attention check, fully responded to the demographic survey, and did not always positively identify each post as referring to substances (*Q-Sub*). Additionally, if this was a sign of unreliable workers, it should be distributed randomly across the demographics. However, this pattern only holds across those who use substances and those who know someone with a substance use disorder and not age, gender, and race/ethnicity.

While we believe these annotations to be useful data, they cause an imbalance in our data set. 65% of the annotations are labeled as stigmatizing, which is much higher than previous studies on toxic language, which tends to find positive labels rare. Therefore, we removed 165 workers who rated at least 95% of their annotations as positive for stigma. This leaves a final data set of 6,147 annotations across 3,802 comments from 400 workers. Their demographic distribution is found in the Final Sample column of Table 2. This final data set has 47% of annotations labeled as stigmatizing.

In Table 3 shows the agreement (Krippendorff's α) for each subgroup (e.g., gender and substance users). The agreement across all groups is $\alpha = 0.12$. We see small differences between Female ($\alpha = 0.14$) and Not Female ($\alpha = 0.13$), as well as African Americans ($\alpha = 0.16$) and Not African Americans ($\alpha = 0.13$). On the other hand, we see larger differences between those with lived experience with substance use. Taken with the results above, we see that those with lived experience are more likely to label a post as stigmatizing but also do not agree on which posts are stigmatizing.

Methods

Our analysis proceeds in three steps: (**RQ1**) training and evaluating the jury learning model, (**RQ2**) evaluating the effect of demographic representation within the jury, and (**RQ3**) identifying linguistic cues associated with stigma. First, we train and evaluate an annotator-level jury learning model to assess how well our model can classify stigma annotations from text, person-level information, and group-level information. Next, we apply the trained jury model to a set of unseen Reddit comments and assess how different jury configurations (e.g., juries consisting of PWUS and those who do not) change the final stigma label. Finally, we identify language associated with labels and examine where different populations agree/disagree on stigmatizing content in order to understand how people perceive stigma.

Jury Learning

Jury learning is a supervised machine learning framework used to predict an individual annotator's label on unseen examples (in our case, stigmatizing content on Reddit) developed by Gordon et al. (2022). Jury learning draws on the notion of juries in the U.S. legal system, specifically through the use of group voting (as opposed to single judge) and the jury selection process. Modeling each annotator in the data set grants practitioners the ability to define the representation of groups of people in the training data. Thus, practitioners can build a jury relevant to the problem at hand from a large pool of annotators. In our example, since the final desired label is whether or not a Reddit post is stigmatizing towards PWUS, we might want the majority of our jurors to know someone in treatment for a SUD. It is easy to imagine similar examples for other toxicity detection-related tasks, such as annotating gender stereotypes, hate speech towards racial minorities, etc.

The learning architecture follows a Deep and Cross Network (DCN), which is a standard recommender system architecture (Wang et al. 2021). Using a movie recommender system as an example, the recommender system will model the movie itself, a person's past viewing history, and who that person is. In more general terms, this architecture models three distinct pieces of information: content (i.e., the movie), group (i.e., who this person is), and person (i.e., the person's viewing history). Applied to this setting, the DCN models the Reddit comment (content), worker demographics (group), and each worker's annotation history (person).

More formally, the DCN consists of an embedding layer,

		Acc	F1	AUC
Baselines	Most Frequent Class	.50	.33	.50
	LIWC†	.55	.54	.59
	LIWC + Dem.†	.59	.58	.63
	Unigrams‡	.58	.57	.61
	Unigrams + Dem.‡	.63	.63	.67
DCN	BERTweet	.59	.58	.59
	BERTweet + Dem.	.64	.64	.64
	BERTweet + Dem. + Ann.	**.69**	**.69**	**.69**

Table 4: Annotation level predictive accuracy. Acc = Accuracy, † logistic regression, ‡ extra trees classifier.

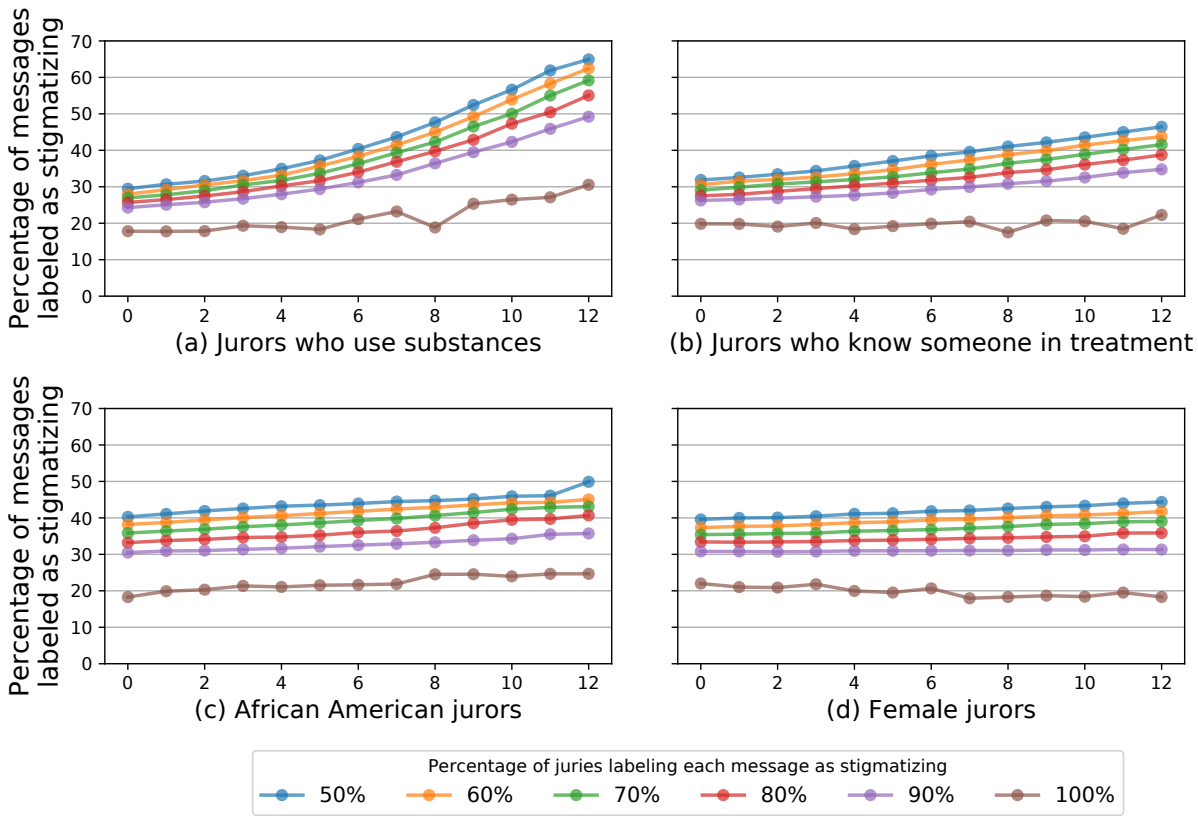

Figure 3: The effect of the jury demographic distribution on the amount of stigmatizing content across the data set. Each dot is the percentage of the 10,000 comments labeled as stigmatizing, where each comment is rated by 10,000 random juries. The color of the line indicates the threshold used to assign the final comment label from the 10,000 jury votes (e.g., a blue line means that a comment is labeled as stigmatizing if at least 50% of the 10,000 juries vote positive stigma). Within each plot, moving from left to right increases the representation of jurors with a given demographic within each of the 10,000 juries (e.g., juries in (d) at 0 are all male and at 12 are all female).

a cross-network, and a deep network. The embedding layer is a concatenation of the content, group, and person embeddings. This is then fed into the cross-network, which models the explicit interactions between the three embedding types through the use of cross layers. The output of the cross-network is then fed into the deep network (a standard feed-forward network) to model implicit interactions.

We emphasize the fact that the end goal for the DCN is a trained architecture that can be used to model each worker in the annotation task (**RQ1**). Thus, the DNC is used to model both the annotators and the data (i.e., the Reddit posts). We further emphasize the fact that during the training process we have not introduced the idea of the "jury" (i.e., jury learning is not limited to DCNs). This is done after we have trained the DCN and applied the model to unseen data to generate stigma predictions from each annotator (**RQ2** and **RQ3**).

RQ1: Automatically Identifying Stigma

We initialize our model and vectorizer with a standard BERTweet model (Nguyen, Vu, and Nguyen 2020; Liu et al. 2019). Using a similar training procedure to Gordon et al. (2022), we train the BERTweet model and the DCN on five epochs of our data, allowing it to alter the underlying BERTweet model. We then freeze the BERTweet model and train for 15 more epochs. We note that Gordon et al. (2022) do an initial fine-tuning step where the BERTweet model is fine-tuned on a large toxicity data set. We chose not to do this initial fine-tuning as the data set is focused on toxicity detection, which may be different from substance use-related stigma. Since our end goal is to understand stigmatizing content, we felt that this fine-tuning process might introduce unintended biases toward general toxicity.

As described in the Data section, the DCN is trained on 80% of the comments in the final annotation data and evaluated on the remaining 20% of comments. We use BERTweet's pooler output as the content embeddings. We include five worker-level demographics (group embedding): continuous age, binary gender (1 for female, 0 otherwise), binary race/ethnicity (1 if African American, 0 otherwise)[2],

[2]Our analysis includes very narrow senses of gender and race/ethnicity and we do not mean to imply that either construct

knowing someone in treatment (1 if the worker knows someone, 0 otherwise), and the (continuous) number of times the worker used substances in the past 30 days. We also include a one-hot encoding of each worker (person embedding). See Appendix for full details and hyperparameter values. Out-of-sample classification accuracy is measured via accuracy, F1, and Area under the ROC Curve (AUC).

Baselines We compare the DCN to two baselines: Linguistic Inquiry and Word Count (LIWC) and unigrams. In all baselines, we consider models trained on (1) text-based features and (2) text-based features plus worker-level demographics (age, gender, race/ethnicity, substance use in the past 30 days, and knowing someone in treatment for a SUD). We also include a simple Most Frequent Class (MFC) classifier. All features are used within either a logistic regression (LIWC) or an Extra Trees Classifier (unigrams). We use the same training and test data as the DCN. See the Appendix for full details and hyperparameter values (which are set via 10-fold cross-validation on the training data).

LIWC LIWC is a dictionary consisting of 73 manually curated categories, including both content and function words, and is one of the most widely used dictionaries in social and psychological sciences (Pennebaker et al. 2015). Example categories include positive and negative emotions, pronouns, verbs, and adjectives.

Unigrams We extract unigrams using a tokenizer designed for social media data (Schwartz et al. 2017). On the training dataset, there are 17,752 unique unigrams. In order to keep the number of features less than the number of observations in the training data (4,761 annotations), we remove rare unigrams: any unigram used by less than .5% of the training data (24 annotations). This results in a total of 1,312 unigrams in the final feature space.

RQ2: Effects of Jury Representation

Here we attempt to answer **RQ2**: Does lived experience with substance use inform how often stigma is perceived? We begin with four binary demographic splits: female/not female, African American/not African American, uses substances/-does not uses substances, knows/does not know someone in treatment. Given a fixed jury size of 12, we consider all possible jury configurations for each of the four demographic splits. For example, a jury with 0 female/12 non-females, 1 female/11 non-females, etc. Then for each of the 10,000 comments in the evaluation data, we create 10,000 random (without replacement) juries for the given configuration (e.g., 10,000 juries with 1 female and 11 non-females). Next, for each of the 10,000 juries, we apply the trained jury model to the jurors to produce 12 stigma ratings (corresponding to the 12 jurors) for the comment. We assign a stigma label of 1 if more than half (7) of the jurors vote that the comment is stigmatizing, and assign 0 otherwise. Thus, for each of the 10,000 comments, we have 10,000 labels, each label produced from the majority vote of the random juries.

is binary.

Next, in order to assign a final stigma label from the 10,000 jury ratings, we consider increasingly stricter thresholds on the percentage of positive stigma votes needed to assign the final label. For example, we begin by assigning a final label to a comment if 50% of the 10,000 juries rate the comment as stigmatizing and increase this threshold up to 100%. We do this for each of the 10,000 comments in the evaluation data set and look at the total percentage of stigmatizing posts across the entire data in order to see if this percentage changes as the demographic representations across the juries change.

RQ3: Stigmatizing Language in the Wild

In order to identify stigmatizing language (**RQ3**), we first assign labels to each of the 10,000 Reddit comments using two separate juries types: a jury where all 12 members have used substances within the last 30 days and a jury where all 12 members have no used substances. At the individual jury level, a stigma vote is assigned if at least 7 members (i.e., the majority) vote that the comment is stigmatizing. As described above, we select 10,000 random juries (for each jury type) and label a comment as stigmatizing if at least 90% of juries vote that the comment is stigmatizing. Thus, in the end, we have two labels for each of the 10,000 comments: one label each from the two jury types.

For the first step of this analysis, we want to know where all juries **Agree** on stigma. Thus, we only consider comments where both jury types agree (e.g., substance-using juries vote *yes* stigma and non-substance-using juries also vote *yes* stigma). We then examine language features (LIWC and unigrams) associated with the binary stigma label. To do this, we perform a single regression for each feature in our feature space (a process called Differential Language Analysis or DLA; Schwartz et al. 2013). In particular, for each language feature, we perform a logistic regression where the independent variable is the relative frequency of a given language feature and the dependent variable is a binary variable set to 0 where both jury types vote *no* stigma and 1 where both jury types vote *yes* stigma. The LIWC category and unigram frequencies are standardized (mean-centered and divided by the standard deviation). Due to the large number of comparisons, we perform a Benjamini-Hochberg False Discovery Rate (FDR) correction and only consider associations significant at a corrected rate of $p < 0.05$ (Benjamini and Hochberg 1995).

For the second step of this analysis, we want to know where people with lived experience with substance use see stigma, but those who do not have the same lived experience do not see stigma. Thus, we only consider comments where substance-using juries vote *yes* stigma and examine where non-substance-using juries **Disagree**. Again, we perform a series of independent logistic regressions (i.e., DLA) using LIWC category and unigram frequencies as the independent variables and a binary dependent variable: 0 where non-substance using juries vote *no* stigma and 1 where non-substance using juries vote *yes* stigma. Again, LIWC category and unigram frequencies are standardized and we apply a Benjamini-Hochberg FDR correction. In both steps, effect sizes are reported as a Cohen's D: the mean difference be-

tween the two groups (the 0 and 1 binary labels) divided by the pooled standard deviation.

Results

Table 4 shows the results of the jury learning process (**RQ1**). Here we see that across all models, adding demographic features increases the predictive accuracy over Reddit comment language alone. We also see that using more sophisticated language features (e.g., BERTweet vs LIWC) increases predictive accuracy. In the end, the DCN (jury learning) using all three feature types (content, person, and group) outperformed all other models. Thus, we can answer **RQ1** in the affirmative: stigma towards PWUS can be automatically identified via machine learning methods.

In Figure 3, we see the results of the jury learning model applied to 10,000 Reddit comments (**RQ2**). Juries with lived experience with substance use (either those who use substances or those who know someone in treatment) tend to label more content as stigmatizing and this increases as their representation within each jury increases. On the other hand, we do not see such pronounced increases across gender or race/ethnicity, both of which are marginalized populations and could be sympathetic to stigma (and thus see stigma where others may not). Here we see slight increases as both African Americans (Figure 3(c)) and females (Figure 3(d)) represent a larger portion of each jury. In reference to **RQ2**, we see that lived experience with substance use increases the *frequency* at which people perceive stigma, and this is not true for other groups.

In Figures 4 and 5, we see language correlated with stigmatizing content (**RQ3**). In Figure 4, the **Agree** label is where both jury types (PWUS and those who do not) agree on stigma/no-stigma, while the **Disagree** label is where juries with PWUS see stigma, but those who do not use substances do not see stigma. As seen in Figure 4 (**Agree** only), stigma is associated with the ANGER, SWEARING, NEGEMO (negative emotions), and SEXUAL categories. The SHEHE category is 3rd person singular pronouns, while PPRON and PRONOUN are general pronoun categories.

Figure 5 gives further context to these results. Here we see references to others ("people", "he", "she", "they", "their"), "addicts" and "addiction", "dealers", and references to children, parents, and schools. Notably, only a single substance is mentioned "meth", which is a highly stigmatized substance (Deen et al. 2021) and the focus of dehumanizing protroyals in the media and in anti-drug campaigns (Habib, Giorgi, and Curtis 2023).

LIWC correlations with the **Disagree** label include INGEST, COGPROC (cognitive processes), and I (first person singular pronouns). Only a few unigrams were associated with the **Disagree** label: "acid", "coke", "i", "weed", and "the". Notably, this includes three specific substance types, whereas the **Agree** results in 5 do not contain many references to substances, other than "meth" which is generally found to be associated with stigmatizing or dehumanizing content (Linnemann and Wall 2013).

While the stigmatizing words associated with **Agree** contain mentions of others, the LIWC category I and the unigram "i" are both associated with **Disgree**. Thus, juries

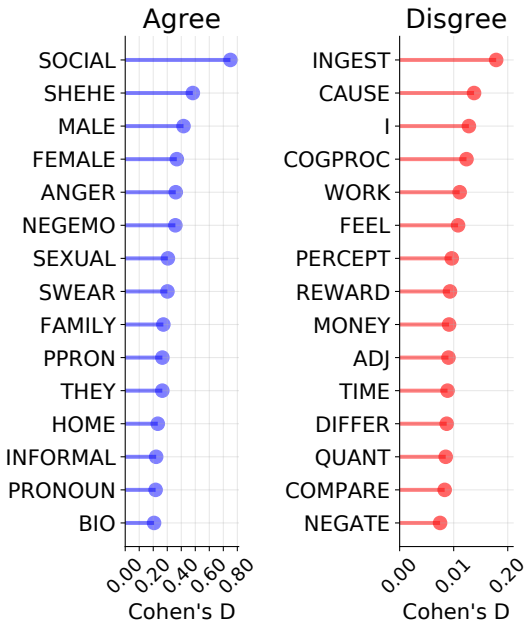

Figure 4: LIWC categories associated with the presence of stigma in comments where both substance using and non-substance using juries agree and disagree (substance using juries labeled as stigma and non-substance using juries did not). All correlations significant at a Benjamini-Hochberg significance level of $p < 0.05$. Note the x-scales of the two plots are different.

which contain PWUS identify stigma in comments which contain self-references, where juries who do not use substances do not find this.

Summarizing these results, in order to answer **RQ3**, we see both similarities and differences between how stigma is perceived between those with lived experience and those without. These two groups agree that negative emotions, swearing, and outgroups (or othering) is indicative of stigma. They also disagree with self-focus and mentions of substances being more stigmatizing for those with lived experience.

Discussion

The results show that (1) those with lived experience perceive more stigma (Figures 1 and 2), (2) those with lived experience do not always agree on what is stigmatizing, and (3) there is overlap between those with lived experience and those without in what words are more like to be stigmatized (Figures 4 and 5). Taken together, there is some agreement across groups on what is stigmatizing, but it is a highly subjective and personal experience. Thus, we believe there are similar takeaways as those from the hate speech/toxicity literature: (1) we need diverse views, (2) we need to include and center those with lived experience, and (3) there is not a "one size fits all" approach to dealing with stigma.

The results also suggest that some substances ("meth")

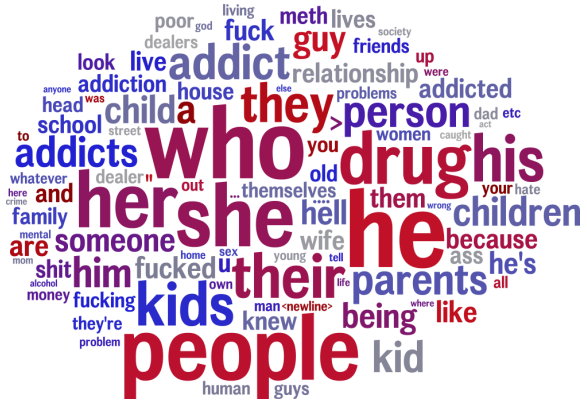

Figure 5: Unigram's associated with the presence of stigma in comments where both substance using and non-substance using juries agree. Size of the word indicates the strength of the correlation (larger Cohen's D); color indicates relative frequency of usage (rare words are grey, frequent words are red, with blue in-between). All correlations significant at a Benjamini-Hochberg significance level of $p < 0.05$. Correlations (Cohen's D) range from 0.07 to 0.31.

may be more stigmatized than others, at least on social media platforms. This differential stigmatization of substances is important to consider given the current political situation in the U.S., where many states are legalizing marijuana while simultaneously defunding harm reduction techniques (such as safe-use sites).

The linguistic correlates of stigma (**RQ3**) dovetail with those found by Chen, Johnny, and Conway (2022). This study also found mentions of family, swearing, negative feelings, and self-focus. Notably missing from our results are mentions of recovery, quitting, and withdrawal, which were all found by Chen, Johnny, and Conway (2022), which might result from their focus on posts about stigmatizing experiences rather than stigmatizing content.

One potential limitation is that the substance use questions and the definition of stigma were presented *before* the actual annotation task. Thus, the annotations could be influenced by the demographic questions and the training material (i.e., priming). We chose to collect the demographic data before the annotation task to maximize the data collected (i.e., if someone does not complete the demographic questions, their annotation data is dropped). This was done since the priming effect could go both ways: substance use questions could influence the annotations and the annotation task could influence the answers to the substance use questions. Similarly, our definition and examples of stigmatizing content (used in the training process before annotations were collected) could also influence how the workers annotated the posts. To minimize this influence, we used examples that we felt were extreme and, thus, likely to be seen as stigmatizing by most people.

Conclusions

Our findings show that people with lived experience with substance use perceive more stigma and respond to different linguistic signals than other marginalized or minority groups (gender or race/ethnicity). Despite this, there is consensus on what types of language lead to stigmatizing content, for example, othering, swearing, and terms like "addict" (though by no means do we wish to imply that stigma is universally agreed upon).

Having a jury-based process for determining stigmatizing language provides a way to gain insights from data by centering diverse groups of workers. This process allows us to conduct a deeper linguistic analysis of messages where those with lived experience with substance use and those who do not differ in their determination of stigma. Our hope is that analysis will reveal concepts and language that seem acceptable to those creating messaging for the substance use audience but are actually harming the recipient. Conversely, there may be messaging that would seem harmful but would instead be impactful.

Broader Perspective

There is increased attention to stigma in both mental health and substance use. This has become especially important in recent years as the U.S. deals with the opioid epidemic, with approximately 1 million deaths since 1999 (Hedegaard et al. 2022). As such, there have been several calls for clinicians and researchers to choose the words they use to describe these stigmatized populations (Volkow, Gordon, and Koob 2021). For example, using the term "poisoning" instead of "overdose" to avoid implications that (1) there is a correct and safe dose and that (2) the substance user knows what a proper dose is and chooses to take more. Similarly, NIDA, previously the National Institute on Drug Abuse, is in the process of changing its name to the National Institute on Drugs and Addiction (National Institutes of Health 2023). Thus, a linguistic analysis of stigma towards PWUS, which centers around people with lived experience, is timely and may help reveal where additional work is needed (Stull et al. 2022).

Ethical Considerations

The study protocol was approved by the Institutional Review Board at the University of Pennsylvania. MTurkers were consented and informed on the nature of the study and what data was being collected. As with all studies using public social media data, none of the Redditors consented to have their comments collected, analyzed, and publicly released. See Chancellor, Baumer, and De Choudhury (2019) for a thorough discussion of the ethical issues (including informed consent) of "humans" in "human-centered machine learning."

We have taken care to anonymize the released data. This includes (1) hashed Mechanical Turk worker-ids, (2) a manual check for identifying information within the Reddit comments, and (3) only releasing Reddit comment-ids instead of releasing the entire comment text. Only releasing comment-ids allows Redditors to delete their posts from the platform,

preventing any future data pulls from collecting their labeled posts (which could be especially problematic given that all data contain substance keywords). Following the recommendations in Gebru et al. (2021), our released data set includes a data sheet with information related to motivation, funding, the collection process, etc. Finally, data will be released using FAIR (Findable, Accessible, Interoperable, Re-usable) guiding principles (Wilkinson et al. 2016).

The Mechanical Turk HITs were written in monolingual English and workers were required to live in the U.S. Similarly, the LIWC dictionaries and BERTweet model used in this study only contain monolingual English. Therefore, the results presented here may not generalize outside the U.S. or to non-English languages or minority populations.

Acknowledgments

We thank Mitchell L. Gordon for sharing his jury learning code. This research was supported in part by the Intramural Research Program of the NIH, National Institute on Drug Abuse (NIDA).

References

Abuse, S.; and Administration, M. H. S. 2021. 2021 National Survey of Drug Use and Health (NSDUH) releases.

Administration, D. E.; et al. 2018. Slang terms and code words: A reference for law enforcement personnel. *DEA Intelligence Report DEAHOU-DIR-022*, 18(2018): 2018–07.

Albadi, N.; Kurdi, M.; and Mishra, S. 2018. Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 69–76. IEEE.

Ashford, R. D.; Brown, A. M.; and Curtis, B. 2019. "Abusing addiction": our language still isn't good enough. *Alcoholism Treatment Quarterly*, 37(2): 257–272.

Ashford, R. D.; Lynch, K.; and Curtis, B. 2018. Technology and social media use among patients enrolled in outpatient addiction treatment programs: cross-sectional survey study. *Journal of medical Internet research*, 20(3): e84.

Basile, V.; Bosco, C.; Fersini, E.; Nozza, D.; Patti, V.; Pardo, F. M. R.; Rosso, P.; and Sanguinetti, M. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th international workshop on semantic evaluation*, 54–63.

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The pushshift reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 830–839.

Benjamini, Y.; and Hochberg, Y. 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.

Byrne, S. K. 2008. Healthcare avoidance: a critical review. *Holistic nursing practice*, 22(5): 280–292.

Chancellor, S.; Baumer, E. P.; and De Choudhury, M. 2019. Who is the "human" in human-centered machine learning: The case of predicting mental health from social media. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–32.

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

Cheng, C.-M.; Chang, C.-C.; Wang, J.-D.; Chang, K.-C.; Ting, S.-Y.; and Lin, C.-Y. 2019. Negative impacts of self-stigma on the quality of life of patients in methadone maintenance treatment: The mediated roles of psychological distress and social functioning. *International journal of environmental research and public health*, 16(7): 1299.

Clement, S.; Schauman, O.; Graham, T.; Maggioni, F.; Evans-Lacko, S.; Bezborodovs, N.; Morgan, C.; Rüsch, N.; Brown, J. S.; and Thornicroft, G. 2015. What is the impact of mental health-related stigma on help-seeking? A systematic review of quantitative and qualitative studies. *Psychological medicine*, 45(1): 11–27.

Corrigan, P. W.; and Watson, A. C. 2002. Understanding the impact of stigma on people with mental illness. *World psychiatry*, 1(1): 16.

Davani, A. M.; Díaz, M.; and Prabhakaran, V. 2022. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics*, 10: 92–110.

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, 512–515.

De Gibert, O.; Perez, N.; García-Pablos, A.; and Cuadros, M. 2018. Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444*.

Deen, H.; Kershaw, S.; Newton, N.; Stapinski, L.; Birrell, L.; Debenham, J.; Champion, K. E.; Kay-Lambkin, F.; Teesson, M.; and Chapman, C. 2021. Stigma, discrimination and crystal methamphetamine ('ice'): Current attitudes in Australia. *International Journal of Drug Policy*, 87: 102982.

Díaz, M.; Johnson, I.; Lazar, A.; Piper, A. M.; and Gergle, D. 2018. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 chi conference on human factors in computing systems*, 1–14.

Fortuna, P.; and Nunes, S. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4): 1–30.

Founta, A. M.; Djouvas, C.; Chatzakou, D.; Leontiadis, I.; Blackburn, J.; Stringhini, G.; Vakali, A.; Sirivianos, M.; and Kourtellis, N. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Golbeck, J.; Ashktorab, Z.; Banjo, R. O.; Berlinger, A.; Bhagwan, S.; Buntain, C.; Cheakalos, P.; Geller, A. A.; Gnanasekaran, R. K.; Gunasekaran, R. R.; et al. 2017. A

large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on web science conference*, 229–233.

Gordon, M. L.; Lam, M. S.; Park, J. S.; Patel, K.; Hancock, J.; Hashimoto, T.; and Bernstein, M. S. 2022. Jury Learning: Integrating Dissenting Voices into Machine Learning Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22. New York, NY, USA: Association for Computing Machinery. ISBN 9781450391573.

Habib, D. R. S.; Giorgi, S.; and Curtis, B. 2023. Role of the media in promoting the dehumanization of people who use drugs. *The American Journal of Drug and Alcohol Abuse*, 1–10.

Hatzenbuehler, M. L.; Phelan, J. C.; and Link, B. G. 2013. Stigma as a fundamental cause of population health inequalities. *American journal of public health*, 103(5): 813–821.

Hedegaard, H.; Miniño, A.; Spencer, M. R.; and Warner, M. 2022. Drug overdose deaths in the United States, 1999–2020. *NCHS data brief*, (428).

Jilka, S.; Odoi, C. M.; van Bilsen, J.; Morris, D.; Erturk, S.; Cummins, N.; Cella, M.; and Wykes, T. 2022. Identifying schizophrenia stigma on Twitter: a proof of principle model using service user supervised machine learning. *Schizophrenia*, 8(1): 1–8.

Kayes, I.; Kourtellis, N.; Quercia, D.; Iamnitchi, A.; and Bonchi, F. 2015. The social world of content abusers in community question answering. In *Proceedings of the 24th international conference on world wide web*, 570–580.

Kelly, J. F. 2004. Toward an addictionary: A proposal for more precise terminology. *Alcoholism Treatment Quarterly*, 22(2): 79–87.

Kelly, J. F.; Wakeman, S. E.; and Saitz, R. 2015. Stop talking 'dirty': clinicians, language, and quality of care for the leading cause of preventable death in the United States. *The American journal of medicine*, 128(1): 8–9.

Li, A.; Huang, X.; Jiao, D.; O'Dea, B.; Zhu, T.; and Christensen, H. 2018. An analysis of stigma and suicide literacy in responses to suicides broadcast on social media. *Asia-Pacific Psychiatry*, 10(1): e12314.

Li, A.; Jiao, D.; Liu, X.; Zhu, T.; et al. 2020. A comparison of the psycholinguistic styles of schizophrenia-related stigma and depression-related stigma on social media: content analysis. *Journal of medical Internet research*, 22(4): e16470.

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

Link, B. G.; and Phelan, J. C. 2001. Conceptualizing stigma. *Annual review of Sociology*, 363–385.

Linnemann, T.; and Wall, T. 2013. 'This is your face on meth': The punitive spectacle of 'white trash'in the rural war on drugs. *Theoretical Criminology*, 17(3): 315–334.

Liu, L.; Cao, Z.; Zhao, P.; Hu, P. J.-H.; Zeng, D. D.; and Luo, Y. 2022. A Deep Learning Approach for Semantic Analysis of COVID-19-Related Stigma on Social Media. *IEEE Transactions on Computational Social Systems*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Luoma, J. B.; Twohig, M. P.; Waltz, T.; Hayes, S. C.; Roget, N.; Padilla, M.; and Fisher, G. 2007. An investigation of stigma in individuals receiving treatment for substance abuse. *Addictive behaviors*, 32(7): 1331–1346.

MacAvaney, S.; Yao, H.-R.; Yang, E.; Russell, K.; Goharian, N.; and Frieder, O. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8): e0221152.

McNeil, K.; Brna, P. M.; and Gordon, K. E. 2012. Epilepsy in the Twitter era: a need to re-tweet the way we think about seizures. *Epilepsy & behavior*, 23(2): 127–130.

National Institutes of Health. 2023. NIDA IC Fact Sheet 2024. https://nida.nih.gov/about-nida/legislative-activities/budget-information/fiscal-year-2024-budget-information-congressional-justification-national-institute-drug-abuse/ic-fact-sheet-2024. Accessed: March 30, 2024.

Nguyen, D. Q.; Vu, T.; and Nguyen, A. T. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 9–14.

Oscar, N.; Fox, P. A.; Croucher, R.; Wernick, R.; Keune, J.; and Hooker, K. 2017. Machine learning, sentiment analysis, and tweets: An examination of Alzheimer's disease stigma on Twitter. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 72(5): 742–751.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830.

Pennebaker, J. W.; Boyd, R. L.; Jordan, K.; and Blackburn, K. 2015. The Development and Psychometric Properties of LIWC2015. Technical report, University of Texas at Austin.

Prabhakaran, V.; Davani, A. M.; and Diaz, M. 2021. On Releasing Annotator-Level Labels and Information in Datasets. In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, 133–138.

Ross, B.; Rist, M.; Carbonell, G.; Cabrera, B.; Kurowsky, N.; and Wojatzki, M. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*.

Rottger, P.; Vidgen, B.; Hovy, D.; and Pierrehumbert, J. 2022. Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 175–190. Seattle, United States: Association for Computational Linguistics.

Roy, A.; Paul, A.; Pirsiavash, H.; and Pan, S. 2017. Automated detection of substance use-related social media posts

based on image and text analysis. In *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*, 772–779. IEEE.

Sap, M.; Card, D.; Gabriel, S.; Choi, Y.; and Smith, N. A. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1668–1678. Florence, Italy: Association for Computational Linguistics.

Sap, M.; Swayamdipta, S.; Vianna, L.; Zhou, X.; Choi, Y.; and Smith, N. A. 2022. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5884–5906. Seattle, United States: Association for Computational Linguistics.

Schwartz, H. A.; Eichstaedt, J. C.; Kern, M. L.; Dziurzynski, L.; Ramones, S. M.; Agrawal, M.; Shah, A.; Kosinski, M.; Stillwell, D.; Seligman, M. E.; et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9): e73791.

Schwartz, H. A.; Giorgi, S.; Sap, M.; Crutchley, P.; Ungar, L.; and Eichstaedt, J. 2017. DLATK: Differential Language Analysis ToolKit. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 55–60. Copenhagen, Denmark: Association for Computational Linguistics.

Sirey, J. A.; Bruce, M. L.; Alexopoulos, G. S.; Perlick, D. A.; Friedman, S. J.; and Meyers, B. S. 2001. Stigma as a barrier to recovery: Perceived stigma and patient-rated severity of illness as predictors of antidepressant drug adherence. *Psychiatric services*, 52(12): 1615–1620.

Stangl, A. L.; Earnshaw, V. A.; Logie, C. H.; van Brakel, W.; C Simbayi, L.; Barré, I.; and Dovidio, J. F. 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): 1–13.

Stull, S. W.; Smith, K. E.; Vest, N. A.; Effinger, D. P.; and Epstein, D. H. 2022. Potential value of the insights and lived experiences of addiction researchers with addiction. *Journal of Addiction Medicine*, 16(2): 135–137.

Uma, A. N.; Fornaciari, T.; Hovy, D.; Paun, S.; Plank, B.; and Poesio, M. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72: 1385–1470.

Van Boekel, L. C.; Brouwers, E. P.; Van Weeghel, J.; and Garretsen, H. F. 2013. Stigma among health professionals towards patients with substance use disorders and its consequences for healthcare delivery: systematic review. *Drug and alcohol dependence*, 131(1-2): 23–35.

Volkow, N. D.; Gordon, J. A.; and Koob, G. F. 2021. Choosing appropriate language to reduce the stigma around mental illness and substance use disorders. *Neuropsychopharmacology*, 46(13): 2230–2232.

Wang, R.; Shivanna, R.; Cheng, D.; Jain, S.; Lin, D.; Hong, L.; and Chi, E. 2021. Dcn v2: Improved deep & cross network and practical lessons for web-scale learning to rank systems. In *Proceedings of the Web Conference 2021*, 1785–1797.

Waseem, Z. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, 138–142.

Waseem, Z.; and Hovy, D. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, 88–93.

Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1): 1–9.

Zhang, Y.; Fan, Y.; Ye, Y.; Li, X.; and Winstanley, E. L. 2018. Utilizing social media to combat opioid addiction epidemic: automatic detection of opioid users from twitter. In *Workshops at the thirty-second AAAI conference on artificial intelligence*.

Paper Checklist

1. For most authors...

   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, the question of automatically identifying stigmatizing content could further marginalize an already vulnerable population, as has been shown in previous work on hate speech (Sap et al. 2019).

   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes.

   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes

   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes

   (e) Did you describe the limitations of your work? Yes, throughout the paper

   (f) Did you discuss any potential negative societal impacts of your work? Yes, in the Introduction.

   (g) Did you discuss any potential misuse of your work? Yes

   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, in the Ethical Considerations section

   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes

2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, in the Appendix
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, we discussed known problems with content moderation.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? Yes
   (b) Did you mention the license of the assets? No, but we acknowledge the research group who privately shared resources.
   (c) Did you include any new assets in the supplemental material or as a URL? Yes
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, in the Ethical Considerations section.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? Yes
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? Yes
6. Additionally, if you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots? Yes, in the Appendix.
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Yes, in the Annotation Task section.
   (d) Did you discuss how data is stored, shared, and deidentified? Yes

Annotation Task Details

Demographics Survey We ask the following demographic questions: (1) What is your age (in years)?; (2) What is your gender identity? Mark all that apply: Male, Female, Genderqueer, Agender, Transgender, Cisgender, A gender not listed (open ended); and (3) With which racial and ethnic group(s) do you identify: American Indian/Alaska Native, Hispanic, Latino/a, or Spanish origin, White, Asian, Middle Eastern or North African, Black or African American, Native Hawaiian or Other Pacific Islander, Another race or ethnicity not listed above (open ended).

Substance Use Related Questions We ask the following questions related to substance use: (1) How many times in the past 30 days have you used an illegal drug or used a prescription medication for nonmedical reasons? For instance, because of the experience or feeling it caused; and (2) Do you know anyone close to you who has ever had treatment (including addiction medication) or attended support groups (ex. AA, NA) for substance or alcohol use disorder: Yes/No.

Quiz We ask workers to answer the following three question quiz: (1) What should NOT be considered drug-related: mentions of a drug store, Insurance plans and prescription coverage, Posts written in a language other than English, All of the above; (2) If you saw the following post, how would you rate it? "I have been taking extra medications because I like the way it makes me feel": Yes, contains drug terms / No, does not contain drug terms; and (3) Which of the following might be considered stigma, or stigmatizing language? Mentions of: prejudice, discrimination, stereotypes, all of the above, none of the above.

Attention Check The following fake Reddit comment was created to be both unambiguously about substances and unambiguously stigmatizing: "The average burglar is

just some idiot, 20 year old crackhead. Stealing from people like me to buy drugs for them and their addict friends. They get what they deserve and belong in jail. At least they can't collect unemployment there." Workers were then asked the same two questions as all other Reddit comments: "Are the drug-related words in this post being used to talk about drugs?" and "Does this post contain stigmatizing language?". This attention check was asked at the halfway point in each task (after 10 of the 20 annotations). Given an incorrect answer (*No* to either of the two questions), the associated annotations were deleted. Despite the fact that the data were not used in the final data set, workers were able to complete the task and were compensated the full amount.

Experimental Parameters

Following Gordon et al. (2022), we use a DCN of 3 cross layers followed by 3 deep layers of size 768 (standard Multi-Layer Perceptrons with ReLU activation) finally feeding into a logit layer with an output of a single real number. We use an Adam optimizer with learning rate = 1×10^{-5}. Models are trained are trained using NVIDIA RTX A6000 GPU.

The baseline models include a logistic regression and extra trees classifier (ETC). Logistic regression parameters: C = 1000000 (for an l_0 penalty approximation), penalty = l2, dual = False, and random_state = 42. ETC parameters: n_jobs = 12, n_estimators = 1000, max_features = sqrt, criterion = gini, min_samples_split = 2, class_weight = balanced_subsample. Unless otherwise specified, all default values are used. Each classifier is implemented using the scikit-learn Python package (Pedregosa et al. 2011) within the DLATK Python package (Schwartz et al. 2017).