Leveraging Contextual Embeddings with Affective, Social, and Behavioral Features for Substance Use Stigma Detection

Anonymous ACL submission

Abstract

Stigma surrounding substance use can result in severe negative consequences for both physical and mental health. To develop effective interventions, identifying situations in which stigma occurs and characterizing its impact are critical. As part of a project to identify facilitators of substance use stigma reduction and to inform the development of interventions for substance use disorder, this study leverages social media data to identify content with a high probability of containing stigma. We create an annotated corpus of 2,214 Reddit posts from subreddits relating to substance use. We train a set of binary classifiers, in which each classifier detects one of three stigma types: Internalized Stigma, Anticipated Stigma, and Enacted Stigma. By combining RoBERTa contextual embeddings and affective, social, and behavioral features, we produce systems that identify instances of substance use stigma for all three stigma types and outperform RoBERTa-only baselines by up to 6.45 macro F1.

1 Introduction

Social stigma surrounding substance use disorders (SUDs) can create negative consequences for health, employment, housing, and relationships (Kulesza et al., 2014). Substance use stigma can prevent individuals from seeking treatment and remaining in treatment programs (Hammarlund et al., 2018), as individuals experiencing stigma may internalize these negative beliefs and feelings, and have diminished self-esteem and recovery capital (Ashford et al., 2019; Bozdağ and Çuhadar, 2022). However, despite the potential harms of substance use stigma, research on its impact on those affected remains limited (Brown, 2011; Livingston et al., 2012; Kulesza et al., 2013, 2017; Smith et al., 2016; Corrigan et al., 2017).

This substance use stigma detection study is one stage of a larger project that seeks to expand our current knowledge of the contexts in which stigma occurs in order to inform the development of future SUD interventions. The current phase of this project involves applying classification methods to identify high-probability instances of substance use stigma in posts extracted from substance use subreddits (discussion forums). To ensure that we capture stigma in the diverse forms in which it occurs, we employ the Stigma Framework (Earnshaw & Chaudoir, 2009), which has been used to conceptualize and measure stigma processes in various contexts, including problematic substance use (Smith et al., 2016) and HIV (Earnshaw & Chaudoir, 2009). In addition to detecting stigmatizing language ("my sister is a hopeless alcoholic"), we also aim to detect reports of stigmatization ("my husband took away the kids and said I'd never get clean"), and reports of the experience of stigma ("I feel so much shame that I can't tell anyone"), which adds an additional layer of difficulty to our task. Our contributions are as follows:

- We propose a hybrid stigma detection model that combines RoBERTa contextual encodings (Liu et al., 2019) with count-based features, allowing the model to leverage affective, social, and behavioral concepts related to substance use stigma.
- We demonstrate that variants of our hybrid model outperform RoBERTa-only baselines and provide an analysis of hybrid model performance.
- We develop and share a set of stigma lexicons informed by stigma theory, along

with our model code, for use in future research involving detection of stigma-related concepts.[1]

2 Related Work

2.1 Stigma Detection

Although a multitude of computational models for the detection of abusive language and hate speech in social media texts has been proposed (Schmidt & Wiegand, 2017; Yin & Zubiaga, 2021), the computational detection of social stigma has been an area less often explored. Whereas hate speech is commonly defined as a communicative act of disparagement of a person or group (Nockleby, 2000), the arguably broader concept of stigma can include, in addition to direct antagonism, more subtle and systematic forms of discrimination and distancing, of both others and the self (Allport et al., 1954; Goffman, 1963). The concept of stigma has been defined differently depending on the circumstances it has been applied to (Link & Phelan, 2001), and so it is not surprising that instead of 'general stigma' detection systems, we see stigma detection systems built toward more specialized purposes. To date, models for the detection of depression stigma (Li et al., 2018), mental health stigma (Lee & Kyung, 2022), stigmatizing language in healthcare discussions (Straton et al., 2020), Alzheimer's Disease stigma (Oscar et al., 2017), and schizophrenia stigma (Jilka et al., 2022) have been proposed.

Li et al. (2018) proposes a system for the detection of depression stigma in Mandarin Chinese Weibo posts. In their data, they find only 6% of the posts contain stigmatizing content; however, when training their model, the authors create a balanced corpus of texts (stigmatizing vs. non-stigmatizing). The researchers test logistic regression, multi-layer perceptron (MLP), support vector machine, and random forest classifiers trained on a simplified Chinese version of Linguistic Inquiry and Word Count (LIWC) features (Pennebaker et al., 2015). The trained models detect stigmatizing posts and also classify each stigma-positive instance as an instance of one of three depression stigma sub-narratives ('unpredictability', 'weakness', or 'false illness'), with the researchers finding best results when using random forest models.

Straton et al. (2020) builds a model for the detection of stigmatizing language in Facebook healthcare discussions around the topic of vaccination. In their annotated corpus of postings from anti-vaccination message walls, they find language stigmatizing government organizations and institutions, and in pro-vaccination message walls, they find language stigmatizing the anti-vaccination movement. Using a balanced dataset, the researchers use term frequency-inverse document frequency (TF-IDF) weighted n-grams and LIWC psychological features to train a variety of classifiers, with a convolutional neural network model resulting in the best performance.

To develop a model for detecting stigmatizing language related to mental health, Lee and Kyung (2022) create a corpus of 240 sentence pairs (stigmatizing and non-stigmatizing), entitled the Mental Health Stigma Corpus. The authors fine-tune a BERT-base model (Devlin et al., 2019) to classify sentences as stigma-positive or stigma-negative and achieve promising results, though the synthetic nature of their dataset may raise questions with regard its ability to generalize to real-world data.

2.2 Imbalanced Learning

In all three of the examples of stigma detection described here (Li et al., 2018; Straton et al., 2020; Lee & Kyung, 2022), the researchers use balanced datasets to both train and evaluate their models. However, in randomly sampled, and even purposively sampled corpora of social media texts, the occurrence of stigma can be relatively rare, with the number of stigma-negative texts (i.e. text containing no evidence of stigma) greatly exceeding the number of stigma-positive ones (Oscar et al., 2017; Li et al., 2018). In such scenarios, the imbalance in data may result in classifiers which perform well for the majority class, but poorly for the minority class (He & Garcia, 2009; Haixiang et al., 2017). This issue can be addressed through imbalanced learning methods such as threshold movement, ensemble learning, and data augmentation.

Threshold Movement. Threshold moving (Song et al., 2014; Zou et al., 2016) can mitigate performance issues related to class imbalance by manipulating the output of the model, setting the decision threshold at a point where performance for the minority class is optimized on a validation set.

Ensemble Learning. Ensemble learning has also been demonstrated as an effective method for dealing with class imbalance (Liu et al., 2009). By

[1] https://anonymous.4open.science/r/stigma_detection-3681

Stigma type	Definition	Synthetic example
Internalized Stigma	The endorsement and application of negative stereotypes about substance users as a group to oneself.	"I'm such a pathetic drunk."
Anticipated Stigma	Expectations that one will experience stereotyping, prejudice, and/or discrimination in the future due to a stigmatized attribute.	"I'll be fired if they find out about my drinking problem."
Enacted Stigma	Past or present experiences of stereotyping, prejudice, and/or discrimination due to a stigmatized attribute.	"My partner left me because of my use."

Table 1: Substance use stigma type definitions adapted from Smith et al. (2016).

combining the outputs of multiple models trained on the same minority class examples, but different subsets of the majority class, training data can be balanced without the loss of majority class information.

Data Augmentation. Beddiar et al. (2021) demonstrate the efficacy of data augmentation (created via back translation) for the task of hate speech detection. Back-translations are produced using machine translation models, which translate from the source language to an intermediate language, and then back to the source language. This results in a paraphrased version of the original text with variations in word choice and other linguistic features; these new perturbed versions of the original data can then be leveraged during training.

2.3 Adding Features to BERT

Based on the effectiveness of BERT contextual embeddings, TF-IDF-weighted n-grams, and LIWC features for the purpose of stigmatizing language detection (Li et al., 2018; Straton et al., 2020; Lee & Kyung, 2022), we choose to experiment with combinations of these resources in our own system. Additionally, given the prevalence of affect types such as sadness, anxiety, and fear in social media posts discussing experiences of SUD recovery (Chen, 2022) and prior literature arguing that emotion regulation can be a factor in stigma coping (Hatzenbuehler et al., 2009; Wang et al., 2018), we experiment with count-based features that include affective, social, and behavioral concepts based on stigma theory, including anxiety, depression, and secretive behavior (Livingston et al., 2012; Kulesza et al., 2013).

Prakash et al. (2020) provide a strategy for combining RoBERTa contextual embeddings with count-based features in order to improve the detection of stance. The researchers create a hybrid model that includes both a RoBERTa encoder, and an MLP which is trained on TF-IDF-weighted n-grams. The authors observed that the MLP component of their system required more epochs of training than typically needed for RoBERTa fine-tuning. To provide adequate training time for the MLP portion of their model, Prakash et al. begin by first pre-training the MLP, and then they combine the pre-trained MLP with RoBERTa before fine-tuning the entire system. The authors' hybrid model outperforms a RoBERTa baseline and achieves state-of-the-art results for stance detection on the RumourEval 2019 dataset (Gorrell et al., 2019).

3 Dataset Creation

Collecting Posts. To create our dataset, approximately 100 thousand English-language Reddit posts authored between January 1, 2013 and December 31, 2019 were collected using Pushshift.io (Baumgartner et al., 2020). Thread-initiating posts were collected from subreddits related to the three substances of interest in the project: alcohol, cannabis, and opioids (e.g., 'r/stopdrinking', 'r/marijuana', and 'r/opiates').

Annotation Process. To select posts for annotation from the harvested data, we utilize keyword sampling, where only posts that match a regular expression containing a keyword list are sampled to increase the probability of sampling stigma-related content. The keyword list includes terms with stigma-related connotations (such as 'shame', 'disappoint', and 'untrustworthy') and terms referring to the actors who may be involved in stigma-related experiences ('family', 'co-worker', 'husband').

Three annotators with expertise in informatics, natural language processing, clinical practice, and public health annotated a total of 2,214 Reddit posts at the span-level for three stigma types based on the Stigma Framework (Earnshaw & Chaudoir, 2009): *Internalized Stigma*, *Anticipated Stigma*, and *Enacted Stigma*. We developed an annotation guide including definitions, synthetic examples, and instructions for identifying and distinguishing these three stigma types based on extant literature (Palamar et al., 2011; Smith et al., 2016). Definitions and examples are presented in Table 1, and a

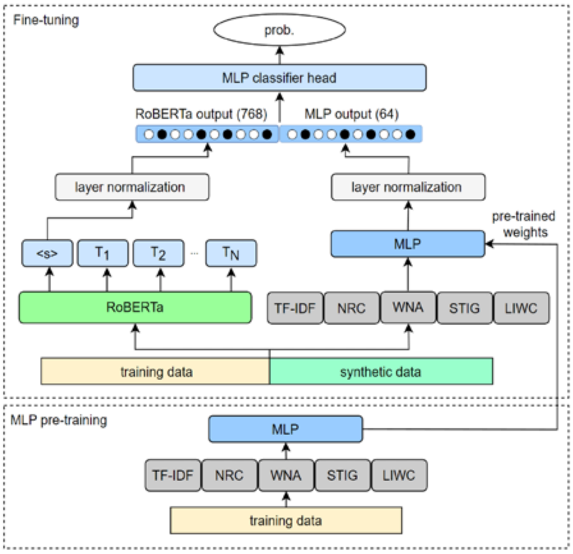

Figure 1: Architecture of the proposed hybrid model.

detailed description of our annotation guidelines is in Appendix C.

Annotators independently identified passages containing stigma in the posts before discussing and reconciling the annotations. Inter-annotator agreement was measured using the F-measure, a measure used as a surrogate for Cohen's Kappa (Cohen, 1960; Hripcsak & Rothschild, 2005).The overall pair-wise F-measure for inter-annotator agreement, prior to reconciliation, varied between 0.67 and 0.71, indicating substantial agreement (Viera et al., 2005).

4 Substance Use Stigma Detection Model

To identify Reddit posts in the harvested data that have a high probability of containing reports and instances of substance use stigma, we create binary classifiers for each stigma type: Internalized Stigma, Anticipated Stigma, and Enacted Stigma.

We utilize a RoBERTa encoder as the main component of our classifier, and also make use of n-gram features, features derived from affective and psychological lexicons, and handcrafted features to provide the model with additional leverage points that are grounded in stigma-related concepts. To integrate RoBERTa embeddings with the additional features, we create a hybrid model (Figure 1), where the first stage is MLP pre-training. The MLP is pre-trained on a concatenated vector of TF-IDF weighted n-grams, features derived from the NRC[2] Emotional Intensity Lexicon (Mohammad, 2018), features derived from Wordnet-Affect (Strapparava & Valitutti, 2004), features generated from the LIWC 2015 lexicon (Pennebaker et al., 2015), and handcrafted substance use stigma features.

Text level	Internalized Stigma	Anticipated Stigma	Enacted Stigma	Total texts
	n / %	n / %	n / %	
Post	764 (34.51%)	420 (18.97%)	361 (16.31%)	2,214
Segment*	1,065 (12.74%)	573 (6.85%)	492 (5.88%)	8,362
Sentence	1,830 (3.96%)	793 (1.72%)	783 (1.69%)	46,215

* Segments are ~600 characters in length.

Table 2: Stigma-positive portion of annotated corpus.

After pre-training is complete, the trained MLP weights are used along with a pre-trained RoBERTa encoder in the fine-tuning process, where the training data is augmented with back-translations. The <s> token output of the RoBERTa encoder and the MLP output are normalized and then concatenated before being passed to an MLP classifier head, which outputs the probability that a given sequence of text contains the current type of substance use stigma.

4.1 Text Segmentation

After tokenizing our corpus, we find that many of our annotated Reddit posts exceed the 512-token RoBERTa input length limit, and thus we opt to chunk posts into segments, and use those segments to train our models. When the trained models make predictions, they first make predictions on individual segments before we map these predictions back to the post level, where, if any segment within a post is predicted as stigma-positive, the entire post is then predicted to be stigma-positive.

Although segmenting posts solves the input limitation issue, this also increases the class imbalance in our dataset. In our annotated corpus, we find that within individual posts, the stigma-positive spans can be infrequent, with multi-paragraph posts sometimes only containing a few stigma-positive words. As a result, when we split the Reddit posts into smaller units (such as sentences), we produce far more negative examples than positive ones, and the portion of stigma-positive texts in our corpus decreases (as shown in Table 2). When splitting posts down to the level of sentences, we see severe class imbalance, with only 1.69% of the data containing Enacted Stigma.

To mitigate class imbalance, we experimented with a variety of segmentation lengths, and found the best performing length to be approximately 600

[2] National Research Council Canada

Feature set	Categories and concepts
NRC Affective Intensity	anger, anticipation, disgust, fear, joy, sadness, surprise, trust, positive, negative
Wordnet-Affect (WNA)	shame, guilt, loneliness, depression, anxiety, anger, confusion, despair, negative-fear, forgiveness, happiness, optimism, sadness
Internalized Stigma (INT)	shame, despair, self-blame, labeling, pejoratives, loss
Anticipated Stigma (ANT)	secrecy, status, awareness, fear, potential consequences, social connections
Enacted Stigma (ENA)	punishment, loss, stigmatizing actions, labeling, pejoratives, trust
LIWC 2015	analytic, clout, authentic, tone, WPS, sixltr, dic, function, pronoun, ppron, i, we, you, shehe, they, ipron, article, prep, auxverb, adverb, conj, negate, verb, adj, compare, interrog, number, quant, affect, posemo, negemo, anx, anger, sad, social, family, friend, female, male, cogproc, insight, cause, discrep, tentat, certain, differ, percept, see, hear, feel, bio, body, health, sexual, ingest, drives, affiliation, achieve, power, reward, risk, focuspast, focuspresent, focusfuture, relativ, motion, space, time, work, leisure, home, money, relig, death, informal, swear, netspeak, assent, nonflu, filler, allpunc, period, comma, colon, semic, qmark, exclam, dash, quote, apostro, parenth, otherp

Table 3: Categories and concepts included in feature sets.

characters. At this length, text segments seem to be short enough to mitigate the amount of confounding information (features unrelated to stigma), but they also remain lengthy enough to keep the imbalance of classes from becoming severe.

To build segments from our post data, we begin by splitting all posts into sentences using NLTK 3.5 (Bird et al., 2009). We then join the resulting sentences in the order they appear in the post until the threshold value of 600 characters in length is reached, after which, a new segment is started. We do not split sentences, and thus segments vary in length. After segmenting texts, labels are assigned to segments by checking for overlap between segment spans and annotation spans. The texts are then pre-processed by removing URLs, hyperlinks, and other html-related text residue.

4.2 Feature Vector Construction

When building input to the MLP component of the classifier, we create the following feature sets:

TF-IDF weighted n-grams (TF-IDF): To create TF-IDF features, we remove English stop words from the text using the NLTK 3.5 package, and then use Scikit-learn 1.8 (Pedregosa et al., 2011) to create TF-IDF weighted n-grams in the range (2, 6) with a dimensionality of 10,000.

LIWC Features: Linguistic, grammatical, and psychological features are generated using LIWC 2015 software (Pennebaker et al., 2015). We remove the 'word count' feature and retain all others, resulting in a 92-dimensional vector.

NRC Affective Intensity Features (NRC): We include NRC features (Mohammad, 2018) to take advantage of the scaled emotional intensity scores that the NRC lexicon provides. We use the NRC Emotional Intensity Lexicon to generate 10-dimensional intensity-scaled affect features (with each dimension corresponding to one of the concepts listed in Table 3). To produce feature vectors, we follow the method of Babanejad et al. (2020), who create 'EAISe' representations (Emotion Affective Intensity with Sentiment Features) for their sarcasm detection model.

Wordnet Affect features (WNA): Wordnet-Affect (Strapparava & Valitutti, 2004), developed based on Wordnet 1.6 (Miller, 1995), enabled us to incorporate finer-grained affect types. Based on literature relating to substance use, stigma, and emotion and an examination of our Reddit corpus, we identified 13 Wordnet-Affect concepts that were relevant to substance use stigma (Table 3) and build lexical sets around each of the 13 Wordnet-Affect concepts using Wordnet. Using these sets, we generate 13-dimensional feature vectors using the same method that we use to build our NRC vectors.

Substance Use Stigma Features (INT / ANT / ENA): We create handcrafted lexicons (identified as 'INT', 'ANT', and 'ENA') to capture specific affective and behavioral concepts related to each stigma type. These lexicons were developed by studying the annotated data, identifying relevant concepts, and iteratively building lexical sets. For Anticipated Stigma, a behavior such as hiding is included in the 'secrecy' concept through keywords such as 'sneak', 'hid', or 'throwaway' (used in mentions of 'throwaway' Reddit accounts created to preserve anonymity). The six concepts included in each feature set is listed here in Table 3, and the complete list of keywords included in each concept is listed in Table 5 of Appendix A. To create 6-dimensional feature vectors, we start with a vector of zeros. We then search text segments for each of the words in our lexical sets. If a lexicon word is present, we add '1' to the concept dimension that the word is associated with.

After building all feature vectors, we separately normalize each set of features, then concatenate them to form a 10,121-dimensional input vector.

4.3 Data Augmentation

During the fine-tuning stage, we augment training data with synthetic data created through back-translation. We use the Google Translate API to translate texts from English to an intermediate language, and then translate back to English, using two languages for backtranslation: Traditional Chinese ('zh-TW') and Japanese ('ja').

4.4 Training

Data Handling. Training sets are sampled from our segment-level data. In development, best results for MLP and hybrid models were found when using a training set with a negative to positive rate of 3:1, and we use this rate to train our final hybrid models. Our validation and test sets are randomly sampled from 10% of the post-level data. After a set of Reddit posts is sampled, the constituent segments are retrieved and used as the evaluation set. After predictions are made on segments, the predictions are then mapped to the post level and evaluation metrics are produced.

Hyperparameters. We train all models on a single Tesla A100 GPU on the Google Colab platform. Training is implemented using Pytorch 1.12 (Paszke et al., 2019) and the Huggingface library (Wolf et al., 2019). We pre-train our MLP for 30 epochs using the AdamW optimizer with a learning rate of 5.e-5 (controlled by a learning rate scheduler) and a batch size of 32. We determine the optimal threshold for positive class F1 after each training epoch using a precision-recall curve on the validation set and the best model is checkpointed based on positive class F1 performance.

During fine-tuning, we fine-tune cased RoBERTa-base (123 million parameters) for 10 epochs with a learning rate of 5.e-5 and batch size of 32. We also experiment with the cased RoBERTa-large encoder (354 million parameters), and when fine-tuning RoBERTa-large, we train for 10 epochs with a learning rate of 7.e-6 and a batch size of 32. Less than 15 minutes of GPU time were required to train a single hybrid model.

Ensemble Training. Our ensemble learning strategy is variance reduction through bootstrap aggregation, or bagging, and we use hard majority voting to produce the final system predictions. For each stigma type, we create an ensemble of five hybrid RoBERTa + MLP models. Each of the models in the ensemble is trained on the same positive examples from the training set, but with a different random sampling of negative examples.

5 Results & Discussion

Table 4 lists the results of post-level stigma detection for the three stigma types. We report the mean macro F1 score of five runs on the same data, using different random seeds. We test variant combinations of features sets to examine which combinations are most effective for each stigma type. As a baseline for comparison to our hybrid models, we list results using RoBERTa-base and RoBERTa-large with a simple classifier head, trained on a balanced training set (via undersampling), and using the same threshold moving method as used in our hybrid model. Additionally, we report MLP evaluation to give some sense of how each feature set might be contributing to performance.

Performance by Stigma Type. Overall, scores for Internalized Stigma are higher than for the other stigma types; Internalized Stigma was the most frequent of the three stigma types in the annotated corpus (making it the stigma type with the greatest number of examples). When performing exploratory feature ranking measures, count-based features had stronger associations (higher chi-square scores) with Internalized Stigma than they did with the other stigma types (Appendix B, Table 6). Keywords such as 'shame' and 'guilt' had strong relationships with Internalized Stigma, which likely benefitted performance.

For Enacted Stigma, overall performance is the weakest of the three stigma types; Enacted Stigma had relatively weak associations with count-based features and the fewest examples. For Enacted Stigma, the highest-ranking features were labels such as 'alcoholic' and 'junkie', which were fairly prevalent across the entire corpus. In our data, we observed that instances of Internalized and Anticipated Stigma frequently focus on a single entity (the post author), with feature rankings for these types showing strong relationships with inward features (n-grams such as 'i ashamed' and 'i lied'). The diffuse nature of Enacted Stigma, involving a more diverse set of actors and behaviors, may be a factor in the difficulty of detecting this stigma type.

In development, we found that RoBERTa-base models consistently outperformed RoBERTa-large for Enacted Stigma. The greater number of parameters in RoBERTa-large seemed to result in overfitting when trained on our limited number of Enacted Stigma examples. Thus, for our final model ensembles, we created a RoBERTa-base hybrid ensemble for Enacted Stigma and a RoBERTa-large

Model	Features	Internalized	Anticipated	Enacted
MLP	TF-IDF	66.06 ± 1.01	58.00 ± 7.21	30.90 ± 2.50
	TF-IDF+NRC	67.73 ± 0.11	58.38 ± 3.52	23.17 ± 1.42
	TF-IDF+NRC+WNA	68.38 ± 0.77	60.04 ± 2.83	30.67 ± 4.79
	TF-IDF+NRC+WNA+STIG	80.03 ± 0.63	68.06 ± 0.96	49.44 ± 2.47
	TF-IDF+NRC+WNA+ STIG+LIWC	72.45 ± 2.77	72.34 ± 2.40	60.64 ± 0.64
RoBERTa-base	-	86.00 ± 1.16	80.04 ± 1.88	70.24 ± 2.10
MLP + RoBERTa-base	TF-IDF	83.46 ± 2.04	82.32 ± 2.07	69.35 ± 1.31
	TF-IDF+NRC	83.64 ± 0.50	80.67 ± 3.61	69.19 ± 1.38
	TF-IDF+NRC+WNA	85.04 ± 1.26	81.83† ± 1.48	70.22 ± 1.33
	TF-IDF+NRC+WNA+STIG	84.49 ± 1.43	84.17† ± 2.60	71.19 ± 2.25
	TF-IDF+NRC+WNA+STIG+LIWC	85.79 ± 1.87	84.23† ± 0.78	69.61 ± 2.80
	TF-IDF+NRC+WNA+ STIG+LIWC, Data aug.	84.51 ± 1.60	81.12 ± 3.17	71.58 ± 1.37
RoBERTa-large	-	85.33 ± 2.29	83.72 ± 2.24	64.60 ± 2.25
MLP + RoBERTa-large	TF-IDF	87.67 ± 1.21	**86.38** ± 1.50	69.51† ± 1.98
	TF-IDF+NRC	88.60 ± 1.53	85.65 ± 2.01	67.12 ± 4.77
	TF-IDF+NRC+WNA	87.17 ± 0.54	85.77 ± 2.32	68.57† ± 2.53
	TF-IDF+NRC+WNA+STIG	87.57 ± 0.61	85.73 ± 1.34	68.46† ± 2.82
	TF-IDF+NRC+WNA+STIG+LIWC	88.34† ± 1.31	84.94 ± 1.16	71.05† ± 0.79
	TF-IDF+NRC+WNA+ STIG+LIWC, Data aug.	87.72 ± 1.39	86.21 ± 1.45	70.58† ± 1.31
Ensemble	TF-IDF+NRC+WNA+ STIG+LIWC, Data aug.	**88.56** ^{L†} ± 0.47	86.30 ^{L} ± 0.36	**72.77** ^{B} ± 2.98

^{L}: MLP + RoBERTa-large
^{B}: MLP + RoBERTa-base

Table 4: Post-level results across models, features, and stigma types. Scores are macro F1 mean values of 5 runs (± std. dev.). Underlined values indicate scores above in-class baseline. † indicates significant improvement over in-class baseline (t-test, p<0.05). **Bold** values indicate the best result for each stigma type.

hybrid ensemble for Internalized Stigma and Anticipated Stigma.

Hybrid Model Performance. For all three stigma types, we found hybrid model variants that significantly outperformed their respective RoBERTa-only baselines, with the largest gain observed for the Enacted Stigma RoBERTa-large model with all feature sets and no data augmentation (+6.45 F1). These results provide evidence that n-gram, affective, social, and behavioral features can be combined with contextual embeddings to improve substance use stigma detection.

For all three stigma types, the 5-model bagging ensembles of fully-featured and data-augmented hybrid models produced slight gains in performance above their single model counterparts.

Impact of Features. Although adding additional feature sets usually led to improvement for MLP models (with some exceptions), we observed less predictable results when adding feature sets to hybrid models. Redundancies in the information encoded by feature set combinations and the information encoded by RoBERTa may have been a factor in the varied performance observed across the hybrid models.

Based on feature rankings (Appendix B), affective features (e.g., WNA_guilt, WNA_shame, and NRC_sadness) appeared to play a greater role in the detection of Internalized Stigma as compared to the other two stigma types. For Anticipated and Enacted Stigma, emotion was still important, but social and behavioral features were also prominent (e.g., ANT_social, ENA_stigmatizing_actions). Anticipated Stigma appeared to include secretive behaviors often involving family, with internal factors such as guilt and shame playing a role, whereas Enacted Stigma involved a more diverse range of interactions with and perceptions of, others. Similar to Straton et al. (2020), we observed that the LIWC categories for emotional tone and clout showed fairly strong relationships with stigma; however, we observed a limited relation to stigma for the remaining 90 LIWC categories.

Impact of Data Augmentation. The use of data augmentation provided limited benefits, with only two of the six fully-featured hybrid models (Enacted Stigma RoBERTa-base and Anticipated Stigma RoBERTa-large) showing an increase in performance when fine-tuning on back translations. In development, we also experimented with the use of back translations during MLP pre-training, and found including back-translations in both pre-training and fine-tuning phases consistently produced weaker models for all stigma types. During fine-tuning, the overall system seemed to benefit more when seeing previously unseen examples across both the RoBERTa and MLP components.

6 Error Analysis

We provide an error analysis for the Enacted Stigma hybrid model ensemble, the weakest performer of the three stigma types, to gain insights into the challenges involved in detecting this form of stigma. We give paraphrased excerpts from our data to demonstrate error types, with features typical of Enacted Stigma texts bolded.

Temporal Errors. We observed that the hybrid Enacted Stigma model produces false positives for texts expressing expectations of future stigmatization, which does not match the temporal requirements of Enacted Stigma annotations (present or past). The following example is representative of this error type:

> *If I come clean, my family will* ***disown me –*** *that isn't even an option. I don't know how I can stop but I just know i have no choice.*

For the RoBERTa-only baseline model, this error type was noticeably less frequent. This may be a limitation of the use of count-based features in the hybrid models, as the model may weighting keywords such as *disown* more heavily than the tense-related syntactic information that has been shown to be encoded by BERT (Jawahar et al., 2019).

Stigmatizing Quitters. During annotation, we observed that individuals abstaining from substance use were pressured by other substance users, often in the context of alcohol use when it is normalized in home or work-related settings. Though this behavior was not annotated as Enacted Stigma, when it appeared in texts, it led to false positive predictions by both the baseline and hybrid models, and is exemplified by the following excerpt:

> *I told my mother I quit drinking and* ***she laughed at me.*** *It really pissed me off. I quit in May and have avoided telling my family because they drink a lot and I didn't want to put up with the questions or* ***judgement****.*

In examples like this, the model seems to leverage features relevant to Enacted Stigma (*she laughed at me*, *judgement*) while failing to learn cues that indicate the mother is an alcohol user critical of another user's abstinence.

Causality. Both the baseline and hybrid Enacted Stigma models were prone to produce false positives for texts where typical features of Enacted Stigma are present, but the cause or motivation behind an action potentially construed as stigmatizing, is unrelated to stereotyping, prejudice, or discrimination. In the following example, the person making a potentially hurtful comment is unaware that the post author is experiencing an SUD:

> *People are starting to figure out something's up, but they don't know what it is. I saw a friend for the first time in a while yesterday, and* ***he said to my face that I looked like shit****, and asked what was wrong.*

Although BERT models have been demonstrated to encode information that can be leveraged to make predictions about causality (Khetan et al., 2022), interpreting the motivations behind the actions described in texts can be a difficult task even for human judgement. We further discuss this issue in our limitations section.

7 Conclusion

This study created an annotated corpus of 2,214 posts from substance recovery subreddits and trained a set of binary classifiers, in which each classifier detected one of three stigma types. By combining contextual embeddings with count-based features, we developed models that identified high-probability instances of substance use stigma and outperformed RoBERTa-only baselines for all three stigma types. Based on our findings, affective, social, and behavioral features appeared to play a significant role in the detection of substance use stigma. We anticipate that the stigma theory-informed constructs represented in our handcrafted lexicons may also be useful to future stigma research in other contexts.

The development of a substance use stigma detection system is the first step toward identifying phenotypes associated with substance use stigma. By using classifiers to identify a large body of substance use stigma narratives in our unseen data, we hope to find possible facilitators and leverage points that could lead to stigma reduction for those experiencing SUDs.

Limitations

Although the purposive sampling used in this study allowed us to develop a sufficient corpus of stigma-positive texts within a reasonable amount of time, our sampling method may also be viewed as one its limitations. By sampling from a limited set of subreddits focused on substance use recovery, we realize that our detection model may not generalize to other types of texts. Additionally, since keyword matching enrichment was used during the sampling process, the distribution of texts in our corpus differs from that of the substance recovery subreddits which they were sampled from. In a more random sampling, it is highly likely that the prevalence of substance use stigma would be lower than the observed prevalence in our enriched sample. When making predictions on random samples from the same recovery subreddits, our models may face performance issues due to the increased imbalance between stigma-positive and stigma-negative texts.

A main goal of the current phase of the project is to identify stigma and descriptions of stigma within narratives. In many of the possible instances of stigma that appear in the narratives, the motivations behind the potentially stigmatizing actions are unclear or unstated. For posts containing sequences such as 'my parents kicked me out of the house', it may be difficult to determine whether the parents' actions are motivated by stigma or by other factors. Causal ambiguity can lead our models to produce errors, and also lead to disagreement among our annotators. For this reason, we choose to cast a wide net during this stage of the project, and instead of attempting to conclusively identify all instances of substance use stigma in our unseen data, we instead attempt to identify instances where stigma is highly probable. In the following phases of the project, we will manually examine individual instances to determine their validity as instances of substance use stigma.

Ethics Statement

Our work has been determined as non-human subject research by the Human Subjects Division at our institution. To reduce the risk of any potential harms to the authors of these sensitive posts, we do not share our annotated dataset publicly. Given that it may be possible to identify post authors based on verbatim quotes, in presentations of our findings, to protect posters' identities, we present synthetic quotations based on the annotated data (Moreno et al., 2013).

References

Allport, G. W., Clark, K., and Pettigrew, T. (1954). *The nature of prejudice*.

Ashford, R. D., Brown, A. M., Canode, B., McDaniel, J., & Curtis, B. (2019). A Mixed-Methods Exploration of the Role and Impact of Stigma and Advocacy on Substance Use Disorder Recovery. *Alcoholism Treatment Quarterly*, *37*(4), 462–480. https://doi.org/10.1080/07347324.2019.1585216

Babanejad, N., Davoudi, H., An, A., & Papagelis, M. (2020). Affective and Contextual Embedding for Sarcasm Detection. *Proceedings of the 28th International Conference on Computational Linguistics*, 225–243. https://doi.org/10.18653/v1/2020.coling-main.20

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). The Pushshift Reddit Dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*, 830–839.

Beddiar, D. R., Jahan, M. S., & Oussalah, M. (2021). Data expansion using back translation and paraphrasing for hate speech detection. *Online Social Networks and Media*, *24*, 100153. https://doi.org/10.1016/j.osnem.2021.100153

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Medi", Inc.".

Bozdağ, N., & Çuhadar, D. (2022). Internalized stigma, self-efficacy and treatment motivation in patients with substance use disorders. *Journal of Substance Use*, *27*(2), 174–180. https://doi.org/10.1080/14659891.2021.1916846

Brown, S. A. (2011). Standardized measures for substance use stigma. *Drug and Alcohol Dependence*, *116*(1), 137–141. https://doi.org/10.1016/j.drugalcdep.2010.12.005

Brown-Johnson, C. G., Cataldo PhD, J. K., Orozco, N., Lisha, N. E., Hickman, N., & Prochaska, J. J. (2015). Validity and Reliability of the Internalized Stigma of Smoking Inventory: An Exploration of Shame, Isolation, and Discrimination in Smokers with Mental Health Diagnoses. *The American Journal on Addictions / American Academy of Psychiatrists in Alcoholism and Addictions*, *24*(5), 410–418. https://doi.org/10.1111/ajad.12215

Chen, A. T., Johnny, S., & Conway, M. (2022). Examining stigma relating to substance use and contextual factors in social media discussions. *Drug and Alcohol Dependence Reports*, *3*, 100061. https://doi.org/10.1016/j.dadr.2022.100061

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Corrigan, P., Schomerus, G., Shuman, V., Kraus, D., Perlick, D., Harnish, A., Kulesza, M., Kane-Willis, K., Qin, S., & Smelson, D. (2017). Developing a research agenda for understanding the stigma of addictions Part I: Lessons from the Mental Health Stigma Literature. *The American Journal on Addictions*, *26*(1), 59–66. https://doi.org/10.1111/ajad.12458

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv:1810.04805 [Cs]*. http://arxiv.org/abs/1810.04805

Earnshaw, V. A., & Chaudoir, S. R. (2009). From Conceptualizing to Measuring HIV Stigma: A Review of HIV Stigma Mechanism Measures. *AIDS and Behavior*, *13*(6), 1160–1177. https://doi.org/10.1007/s10461-009-9593-3

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, *73*, 220–239. https://doi.org/10.1016/j.eswa.2016.12.035

Hammarlund, R., Crapanzano, K. A., Luce, L., Mulligan, L., & Ward, K. M. (2018). Review of the effects of self-stigma and perceived social stigma on the treatment-seeking decisions of individuals with drug- and alcohol-use disorders. *Substance Abuse and Rehabilitation*, *9*, 115–136. https://doi.org/10.2147/SAR.S183256

He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Hripcsak, G., & Rothschild, A. S. (2005). Agreement, the F-Measure, and Reliability in Information Retrieval. *Journal of the American Medical Informatics Association*, *12*(3), 296–298. https://doi.org/10.1197/jamia.M1733

Goffman, E. (1963). *Stigma: Notes on the Management of Spoiled Identity*. New York: Touchstone.

Gorrell, G., Kochkina, E., Liakata, M., Aker, A., Zubiaga, A., Bontcheva, K., & Derczynski, L. (2019). SemEval-2019 Task 7: RumourEval, Determining Rumour Veracity and Support for Rumours. *Proceedings of the 13th International Workshop on Semantic Evaluation*, 845–854. https://doi.org/10.18653/v1/S19-2147

Hatzenbuehler, M. L., Nolen-Hoeksema, S., & Dovidio, J. (2009). How Does Stigma “Get Under the Skin”?: The Mediating Role of Emotion Regulation. *Psychological Science*, *20*(10), 1282–1289. https://doi.org/10.1111/j.1467-9280.2009.02441.x

Jawahar, G., Sagot, B., & Seddah, D. (2019). What Does BERT Learn about the Structure of Language? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3651–3657. https://doi.org/10.18653/v1/P19-1356

Jilka, S., Odoi, C. M., van Bilsen, J., Morris, D., Erturk, S., Cummins, N., Cella, M., & Wykes, T. (2022). Identifying schizophrenia stigma on Twitter: A proof of principle model using service user supervised machine learning. *Schizophrenia*, *8*(1), 1–8. https://doi.org/10.1038/s41537-021-00197-6

Khetan, V., Ramnani, R., Anand, M., Sengupta, S., & Fano, A. E. (2022). Causal BERT: Language Models for Causality Detection Between Events Expressed in Text. In K. Arai (Ed.), *Intelligent Computing* (pp. 965–980). Springer International Publishing. https://doi.org/10.1007/978-3-030-80119-9_64

Kulesza, M., Larimer, M. E., & Rao, D. (2013). Substance Use Related Stigma: What we Know and the Way Forward. *Addictive Behaviors, Therapy & Rehabilitation*, *2013*. https://doi.org/10.4172/2324-9005.1000106

Kulesza, M., Ramsey, S., Brown, R., & Larimer, M. (2014). Stigma among Individuals with Substance Use Disorders: Does it Predict Substance Use, and Does it Diminish with Treatment? *Journal of Addictive Behaviors, Therapy & Rehabilitation*, *3*(1), 1000115. https://doi.org/10.4172/2324-9005.1000115

Kulesza, M., Watkins, K. E., Ober, A. J., Osilla, K. C., & Ewing, B. (2017). Internalized stigma as an independent risk factor for substance use problems among primary care patients: Rationale and preliminary support. *Drug and Alcohol Dependence*, *180*, 52–55. https://doi.org/10.1016/j.drugalcdep.2017.08.002

Li, A., Jiao, D., & Zhu, T. (2018). Detecting depression stigma on social media: A linguistic analysis. *Journal of Affective Disorders*, *232*, 358–362. https://doi.org/10.1016/j.jad.2018.02.087

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

Liu, X.-Y., Wu, J., & Zhou, Z.-H. (2009). Exploratory Undersampling for Class-Imbalance Learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *39*(2), 539–550. https://doi.org/10.1109/TSMCB.2008.2007853

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (arXiv:1907.11692). arXiv. https://doi.org/10.48550/arXiv.1907.11692

Livingston, J. D., Milne, T., Fang, M. L., & Amari, E. (2012). The effectiveness of interventions for reducing stigma related to substance use disorders: A systematic review. *Addiction (Abingdon, England)*, *107*(1), 39–50. https://doi.org/10.1111/j.1360-0443.2011.03601.x

Lee, M. H., & Kyung, R. (2022). Mental Health Stigma and Natural Language Processing: Two Enigmas Through the Lens of a Limited Corpus. *2022 IEEE World AI IoT Congress (AIIoT)*, 688–691. https://doi.org/10.1109/AIIoT54504.2022.9817362

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*(11), 39–41. https://doi.org/10.1145/219717.219748

Mohammad, S. (2018, May). Word Affect Intensities. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. LREC 2018, Miyazaki, Japan. https://aclanthology.org/L18-1027

Moreno, M. A., Goniu, N., Moreno, P. S., & Diekema, D. (2013). Ethics of Social Media Research: Common Concerns and Practical Considerations. *Cyberpsychology, Behavior, and Social Networking*, *16*(9), 708–713. https://doi.org/10.1089/cyber.2012.0334

Nockleby, J. T., Levy, L. W., Karst, K. L., & Mahoney, D. J. (2000). Encyclopedia of the American constitution. *Detroit, MI: Macmillan Reference*, *3*(2), 1277-1279

Oscar, N., Fox, P. A., Croucher, R., Wernick, R., Keune, J., & Hooker, K. (2017). Machine Learning, Sentiment Analysis, and Tweets: An Examination of Alzheimer's Disease Stigma on Twitter. *The Journals of Gerontology: Series B*, *72*(5), 742–751. https://doi.org/10.1093/geronb/gbx014

Palamar, J. J., Kiang, M. V., & Halkitis, P. N. (2011). Development and Psychometric Evaluation of Scales that Assess Stigma Associated With Illicit Drug Users. *Substance Use & Misuse*, *46*(12), 1457–1467. https://doi.org/10.3109/10826084.2011.596606

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32* (pp. 8024–8035). Curran Associates, Inc. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

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., & 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., & Blackburn, K. (2015). *The Development and Psychometric Properties of LIWC2015*. https://repositories.lib.utexas.edu/handle/2152/31333

Prakash, A., & Tayyar Madabushi, H. (2020). Incorporating Count-Based Features into Pre-Trained Models for Improved Stance Detection. *Proceedings of the 3rd NLP4IF Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, 22–32. https://aclanthology.org/2020.nlp4if-1.3

Schmidt, A., & Wiegand, M. (2017). A Survey on Hate Speech Detection using Natural Language Processing. *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, 1–10. https://doi.org/10.18653/v1/W17-1101

Smith, L. R., Earnshaw, V. A., Copenhaver, M. M., & Cunningham, C. O. (2016). Substance use stigma: Reliability and validity of a theory-based scale for substance-using populations. *Drug and Alcohol Dependence*, *162*, 34–43. https://doi.org/10.1016/j.drugalcdep.2016.02.019

Song, B., Zhang, G., Zhu, W., & Liang, Z. (2014). ROC operating point selection for classification of imbalanced data with application to computer-aided polyp detection in CT colonography. *International Journal of Computer Assisted Radiology and Surgery*, *9*(1), 79–89. https://doi.org/10.1007/s11548-013-0913-8

Strapparava, C., & Valitutti, A. (2004). WordNet-Affect: An Affective Extension of WordNet. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, 1083–1086.

Straton, N., Jang, H., & Ng, R. (2020). Stigma Annotation Scheme and Stigmatized Language Detection in Health-Care Discussions on Social Media. *Proceedings of the 12th Language Resources and Evaluation Conference*, 1178–1190. https://aclanthology.org/2020.lrec-1.148

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine*, *37*(5), 360–363.

Wang, K., Burton, C. L., & Pachankis, J. E. (2018). Depression and Substance Use: Towards the Development of an Emotion Regulation Model of Stigma Coping. *Substance Use & Misuse*, *53*(5), 859–866. https://doi.org/10.1080/10826084.2017.1391011

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., … Rush, A. M. (2019). *HuggingFace's Transformers: State-of-the-art Natural Language Processing*. https://doi.org/10.48550/arXiv.1910.03771

Yin, W., & Zubiaga, A. (2021). Towards generalisable hate speech detection: A review on obstacles and solutions. *PeerJ Computer Science*, *7*, e598. https://doi.org/10.7717/peerj-cs.598

Zou, Q., Xie, S., Lin, Z., Wu, M., & Ju, Y. (2016). Finding the Best Classification Threshold in Imbalanced Classification. *Big Data Research*, *5*, 2–8. https://doi.org/10.1016/j.bdr.2015.12.001

Stigma type	Concept	Keywords
Internalized Stigma (INT)	shame	'shame', 'guilt', 'regret', 'underachiev', 'embarrassed', 'embarrassment', 'loathing', 'embarrassing', 'self respect', 'remorse', 'humiliated', 'humiliation', 'burden',
	despair	'despair', 'hopeless', 'disappointed', 'regret', 'wast', 'tired', 'miser', 'suicid', 'defeated', 'depressed'
	self-blame	'deserve', 'blame', 'self', 'fault', 'fail', 'relapse', 'lack', 'incapable', 'hate myself'
	labels	'stoner', 'addict', 'junkie', 'alcoholic', 'drunk', 'loser', 'zombie', 'pothead', 'crackhead', 'druggie', 'failure', 'asshole', 'idiot', 'fool', 'trash', 'monster', 'degenerate'
	pejoratives	'disgust', 'lazy', 'stupid', 'annoying', 'weak', 'selfish', 'piece of shit', 'inept', 'worthless', 'disappointment', 'pathetic', 'embarrassment', 'awful', 'irresponsible', 'liar', 'horrible', 'foolish', 'shitty', 'unproductive'
	loss	'loss', 'lost', 'lose', 'losing', 'cost', 'ruin', 'ruined', 'wasted'
Anticipated Stigma (ANT)	secrecy	'secret', 'secrecy', 'sneak', 'snuck', 'hid', 'throwaway', ' irl', 'suspect', 'find out', 'finding out', 'finds out', 'admit', 'to tell', ‘t tell’, ‘t talk’, ‘never tell’, ' lie', 'lying', 'truth', 'caught', 'decept', 'outed', 'something is up', 'steal', 'stole', 'pretend', 'suspicious', 'confide', 'transparent', 'come clean', 'coming clean', 'double life', 'account', 'plain sight', 'trackmarks', 'track marks', 'stash', 'excuse', 'red handed', 'red-handed', 'honest'
	status	'judg', 'respect', 'trust', 'shame', 'shun', 'embarras', 'stigma', 'trust', 'reputation', 'taint', 'credibility', 'think less of', 'treated like', 'disappoint', 'believe me', 'intoler', 'labeled'
	awareness	‘everyone knows’, ‘everyone knew’, ‘t know’, ‘never knew’, ‘unaware’, ‘no idea’, ‘any idea’, ‘no one knows’, ‘no clue’, ‘oblivious’, ‘t notice’, ‘t told’
	fear	'fear', 'freaking', 'worr', 'scare', 'afraid', 'eating me up', 'terrified', 'terrifies', 'paranoid', 'anxiety', 'panic', 'tired', 'nervous', 'uncomfortable'
	potential consequences	'to face', 'lose', 'losing', 'cost', 'ruin', 'destroy', 'leave me', 'cut me out', 'ditch', 'leav', 'disown', 'give up on', 'dump', 'distance', 'break up', 'divorce', 'fire', 'alienat'
	social	'family', 'children', 'friend', 'parents', 'dad', 'father', 'mom', 'mum', 'mother', 'husband', 'wife', 'brother', 'sister', 'relationship', 'doctor', 'nurse', 'social circle', 'partner', 'girlfriend', 'boyfriend', 'worker', 'in-law', 'loved ones', 'psychiatrist', 'therapist', 'people I love', 'ones I love',
Enacted Stigma (ENA)	punishment	'legal', 'arrest', 'police', 'ground', 'caught', 'kicked out', 'evict', 'trouble', 'consequence', 'DUI', 'jail', 'prison', 'parole', 'officer', 'charged', 'bust', 'court', 'fired'
	loss	'lost', 'lose', 'losing','cost', 'ruin', 'ruined', 'gone', 'left me', 'cut me out', 'ditch', 'leaving', 'disown', 'gave up', 'dump', 'distance', 'broke up', 'break up', 'nothing to do with', 'anything to do with', 'divorce', 'not welcome', 'estranged'
	stigmatizing-actions	'called', 'blame', 'control', 'mock', 'make fun', 'made fun', 'made me', 'scared', 'ruined', 'judge', 'react', 'laughed', 'looked down on', 'freaked out', 'shun', 'shame', 'assume', 'confront', 'disrespect', 'stigma', 'bully', 'condemn', 'berate', 'insult', 'pigeonhol', 'treated like', 'treat like', 'ridicule', 'spit on', 'spat on', 'wind up like', 'end up like', 'pressure'
	labels	'stoner', 'addict', 'junkie', 'alcoholic', 'drunk', 'loser', 'stereotype', 'zombie', 'thief', 'pothead', 'crackhead', 'druggie', 'criminal', 'pill head', 'fiend', 'tweeker', 'failure', 'asshole', 'idiot', 'scum'
	pejoratives	'disgust', 'lazy', 'stupid', 'annoying', 'weak', 'negative', 'selfish', 'hopeless', 'piece of shit', 'nasty', 'inept', 'crazy', 'worthless', 'disappointment', 'annoying', 'pathetic', 'embarrassment', 'awful', 'irresponsible', 'liar', 'horrible'
	trust	'trust', 'respect', 'insecure', 'disappoint', 'excuse', 'believe', 'lie', 'lying', 'accuse', 'confront', 'cold shoulder', 'suspicious', 'truth', 'found out', 'privacy', 'apologize', 'faith', 'genuine'

Table 5: Keywords included in substance use stigma feature lexicons

A Substance Use Stigma Keywords

Table 5 lists the specific keywords included in handcrafted features sets for Internalized Stigma (INT), Anticipated Stigma (ANT), and Enacted Stigma (ENA). We create handcrafted lexicons for each stigma type by studying the annotated data for each stigma type, identifying relevant concepts, and iteratively building lexical sets.

	Internalized Stigma		Anticipated Stigma		Enacted Stigma	
Rank	**Feature**	**χ2**	**Feature**	**χ2**	**Feature**	**χ2**
1	INT_shame	613.969	ANT_secrecy	168.195	ENA_labels	44.124
2	WNA_guilt	248.967	i lied	33.678	ENA_stigmatizing_actions	32.575
3	WNA_shame	187.890	ANT_social	32.748	LIWC_Tone	27.341
4	INT_self_blame	120.265	i hid	27.649	ENA_trust	21.451
5	LIWC_Tone	109.170	i hiding	26.505	give shit	17.450
6	INT_pejoratives	58.765	ANT_status	24.107	ENA_loss	15.284
7	INT_despair	57.126	LIWC_Tone	24.008	i arrested	14.669
8	INT_labels	54.435	i hide	19.0178	my sister	13.286
9	NRC_negative	52.972	secret i	17.834	low key	11.589
10	LIWC_Clout	49.874	hide family	16.513	my husband	10.736
11	INT_loss	43.394	i tell anyone	14.703	treated like	10.519
12	shame guilt	37.727	one knows	13.688	empty bottles	10.507
13	i ashamed	37.511	hit pen	13.439	drunk last	8.696
14	ashamed i	32.121	i tell	13.206	self respect	8.521
15	guilt shame	28.220	track marks	13.189	ENA_punishment	8.391
16	the shame	25.744	WNA_shame	12.608	so said	8.383
17	feel ashamed	24.252	lied i	12.473	NRC_negative	8.274
18	i embarrassed	24.072	knows i	12.412	lied i	8.182
19	NRC_sadness	23.350	ANT_fear	11.039	my dad	8.142
20	shame i	22.597	WNA_guilt	10.983	drug addict	7.558
21	like failure	22.303	tell family	10.901	i driven	7.511
22	self loathing	22.176	tell anyone	9.470	junkie i	7.478
23	lot shame	19.243	family friends	9.470	what helped	7.463
24	i feel	19.052	i want anyone	9.444	even talk	7.193
25	i hate	18.148	taper so	9.172	home parents	7.112
26	i ashamed i	15.401	tell parents	8.718	because i	7.090
27	i feel like failure	15.369	embarrassed tell	8.614	became addicted	6.881
28	failure i	13.561	i blew	8.323	drug addicts	6.677
29	feel worthless	13.484	coming clean	8.302	think going	6.565
30	loser i	13.446	ANT_awareness	8.133	drunk last night	6.431

NRC Affective Intensity Lexicon feature
Wordnet-Affect feature
Substance use stigma feature (INT/ANT/ENA)
LIWC 2015 feature

Table 6: Top 30 chi-square feature ranking for TF-IDF weighted n-grams, NRC, WNA, INT, ANT, ENA, and LIWC features. Features names (other than n-grams) include a prefix (e.g., 'LIWC_') and color code to indicate feature set membership. All scores are significant at $p < .01$.

B Feature Ranking

We perform exploratory feature ranking for all features included in the input to the MLP portion of our hybrid model, including TF-IDF weighted n-grams, NRC features, Wordnet-Affect features, LIWC 2015 features, and the handcrafted stigma concepts for each stigma type. We use the training set to explore the strength of association between each feature and its relevant stigma type using the chi-square measure. The feature selection tools of the Scikit-learn package were used to implement this experiment (Pedregosa et al., 2011). Results are listed in Table 6, with all scores being significant at $p < .01$.[3]

C Annotation Guidelines

The following is paraphrased and condensed from the guide used by the annotators. In addition to the textual content below, the annotators were also provided scale instruments informed by stigma theory, which assisted them to identify and distinguish the three stigma types (Palamar et al., 2011; Brown-Johnson et al., 2015; Smith et al., 2016; Kulesza et al., 2017).

Guide text:

[3] Note that this experiment does not directly measure the contribution of each feature to model performance; however, it does provide an indication of the strength of the relationships between features and each of the three stigma types.

We are annotating probable occurrences of three different types of stigma: *Internalized, Anticipated, and Enacted Stigma*. These probable occurrences will serve as training data to train classifiers to predict instances of stigma in a larger dataset. We will then perform content analysis of the instances that the classifier identifies to identify leverage points for future interventions.

Because we want to be able to make more nuanced differentiations of stigma through manual review later, we are employing coarser definitions (probable as opposed to certain stigma). This will enable us to later distinguish between human reactions in difficult circumstances, and stigma.

Annotate probable occurrences of stigma:

- Annotate at span level.
- Annotate as much of the text as needed to capture the instance of stigma. This could be part of a sentence, one sentence, or multiple sentences.

Please review the definitions below:

Enacted Stigma: Past or present experiences of stereotyping, prejudice, and/or dis-crimination due to a stigmatized attribute.

Example: *My husband called me an addict and said I'd never become clean, so he was taking the kids away.*

- Annotate this even if the causal attribution is not clear.
- Do not annotate instances in which those who engage in substance use treat someone who has quit or is trying to, in a negative way.
- Annotate situations in which stigma is expressed having to do with a substance that is used to quit the target substance in question (alcohol, cannabis, opioids). An example would be when a person criticizes the use of suboxone for quitting.
- Annotate instances in which actual substance use is not mentioned, but someone mentions enacted stigma relating to persons who uses substances more generally.
- Take what the person says as at face value (accept what they perceive as reality, as opposed to trying to assess whether things are really as they say they are).
- Annotate situations in which people experience legal consequences due to substance use, such as receiving a DUI or being arrested.

Anticipated Stigma: Expectations that one will experience stereotyping, prejudice, and/or discrimination in the future due to a stigmatized attribute.

Example: *Though I don't know if they know, I wonder if my co-workers talk about me and my "problem".*

- This would include: perceptions of society towards substance use, situations in which someone is hiding their habit, being secretive, deceiving others or lying about their habit, and stealing.
- If a person says that they think that negative consequences would occur due to their substance use being found out, it could be considered Anticipated Stigma.
- Annotate this even if the causal attribution is not clear.
- Annotate instances in which someone is surprised that they were not treated badly due to their substance use, or instances in which someone anticipates that they will be treated with prejudice, even if that turns out to not be the case (e.g., a child expects that the parent will turn them out of the house, but the parent says that they understand and they will support them through their situation).

Internalized Stigma: The endorsement and application of negative stereotypes about sub-stance users as a group to oneself.

Example: *I'm a stoner. I am an awful person...*

- This may involve self-incrimination in relation to substance use.
- This may also be manifest as hopelessness and/or weakness (however, hopelessness and/or weakness on their own, is not enough to constitute Internalized Stigma).
- We might consider a concept such as "hopelessness" carrying more weight if it

is in the title. (For example, if hopelessness comes up in the title, we can annotate it as an indication of self-stigma due to its being in a substance use-related discussion forum.)

Do annotate:

- Examples in which the poster is not the main actor involved in the stigma.

Do not annotate:

- Fictional stories or articles (identify stories of stigma that are actually true).
- Dreams.
- Predictions or hypothetical situations.
- Do not annotate across paragraph breaks.
- Do not annotate stigma due to reasons other than substance use, unless they are mixed with substance use stigma. For example: do not annotate the expression of depression on its own, disconnected to feeling badly about one's use of substances.

Other notes:

- Recognizing that you have a problem is not necessarily indicative of stigma (there is a difference between helpful self-reflection and self-stigmatization).
- Distinguish between stigma and substance use. A recurrence of substance use is not an example of Internalized Stigma.
- When we see examples of stigma in the past, code them as stigma, except when the person says they no longer experience it. For example, if the person says they no longer feel shame or they no longer feel worthless, then do not code it as Internalized Stigma.
- If a person says that the substance makes them lazy or results in negative consequences (such as getting into accidents), it is not necessarily indicative of stigma. We will annotate it as stigma if the passage seems to convey a logic where the person seems to feel that people who use that substance are lazy, and since they themselves use the substance, then they are lazy.
- Humiliation: Humiliation can be internal or external. As such, when you encounter an instance of humiliation, think about whether the person is feeling humiliated (likely Internalized Stigma), or whether someone said something to them in response to something that they did (likely Enacted Stigma).