BMC Digital Health

RESEARCH

Open Access

A theory-informed deep learning approach to extracting and characterizing substance use-related stigma in social media

David Roesler[1], Shana Johnny[2], Mike Conway[3*] and Annie T. Chen[1*]

Abstract

Background Stigma surrounding substance use can result in severe consequences for physical and mental health. Identifying situations in which stigma occurs and characterizing its impact could be a critical step toward improving outcomes for individuals experiencing stigma. As part of a larger research project with the goal of informing the development of interventions for substance use disorder, this study leverages natural language processing methods and a theory-informed approach to identify and characterize manifestations of substance use stigma in social media data.

Methods We harvested social media data, creating an annotated corpus of 2,214 Reddit posts from subreddits relating to substance use. We trained a set of binary classifiers; each classifier detected one of three stigma types: Internalized Stigma, Anticipated Stigma, and Enacted Stigma, from the Stigma Framework. We evaluated hybrid models that combine contextual embeddings with features derived from extant lexicons and handcrafted lexicons based on stigma theory, and assessed the performance of these models. Then, using the trained and evaluated classifiers, we performed a mixed-methods analysis to quantify the presence and type of stigma in a corpus of 161,448 unprocessed posts derived from subreddits relating to substance use.

Results For all stigma types, we identified hybrid models (RoBERTa combined with handcrafted stigma features) that significantly outperformed RoBERTa-only baselines. In the model's predictions on our unseen data, we observed that Internalized Stigma was the most prevalent stigma type for alcohol and cannabis, but in the case of opioids, Anticipated Stigma was the most frequent. Feature analysis indicated that language conveying Internalized Stigma was predominantly characterized by emotional content, with a focus on shame, self-blame, and despair. In contrast, Enacted Stigma and Anticipated involved a complex interplay of emotional, social, and behavioral features.

Conclusion Our main contributions are demonstrating a theory-based approach to extracting and comparing different types of stigma in a social media dataset, and employing patterns in word usage to explore and characterize its manifestations. The insights from this study highlight the need to consider the impacts of stigma differently by mechanism (internalized, anticipated, and enacted), and enhance our current understandings of how each stigma mechanism manifests within language in particular cognitive, emotional, social, and behavioral aspects.

*Correspondence:
Mike Conway
michaelambroseconway@gmail.com
Annie T. Chen
atchen@uw.edu
Full list of author information is available at the end of the article

Keywords Substance use, Stigma, Social media, Machine learning, Information extraction

Background

Persons with substance use disorders (SUDs) can experience stigma in various forms, including stereotypes, prejudice, and discrimination, and this stigma can have far-ranging consequences for their health, employment, housing, and relationships [1]. Individuals experiencing stigma may internalize these negative beliefs and feelings, have diminished self-esteem and recovery capital [2, 3], and be reluctant to seek treatment [4].

Interventions focused on stigma reduction in the context of substance use have been limited, and these have tended to focus on structural stigma (e.g., education of professionals that work with persons with SUDs) as opposed to social or self-stigma [5]. There is also awareness of the bias in words used to describe SUDs, and the need to consider word choice [6, 7]. However, despite the potential harms of substance use stigma, our knowledge of how different types of stigma affect persons within the context of SUDs remains limited [5, 8–12].

In this article, we demonstrate a stigma theory-informed deep learning approach to the task of identifying examples of substance use stigma in a large dataset. To ensure that we capture stigma in the diverse forms in which it occurs, we employ the Stigma Framework [13], which defines three stigma mechanisms for those who experience stigma: *Internalized Stigma*, *Anticipated Stigma*, and *Enacted Stigma*. The Stigma Framework has been used to characterize stigma processes in various health-related contexts, including problematic substance use [11] and HIV [13], and extant literature has sought to develop instruments to assess the experience of these three types of stigma [11]. To our knowledge, however, prior work has not explored how the three stigma mechanisms are conveyed by the language used in social media. We examine stigma as expressed in social media for two main reasons: 1) previous literature has shown that stigma relating to mental health is endemic in social media [14, 15]; and 2) social media can serve an important role in understanding and promoting public health [16, 17].

This current study aims to answer the research question: How do the three stigma mechanisms in the Stigma Framework manifest differently in terms of distribution and nature in social media? We take the following approach:

- We develop classifiers to identify three stigma mechanisms in an annotated social media dataset and evaluate the performance of these classifiers.
- To gain a deeper understanding of the prevalence of the three stigma mechanisms in social media at large, we analyze how each stigma mechanism is distributed in the predictions made by the classifiers on the unseen portion of our data.
- To better understand the linguistic expression of the different stigma mechanisms in social media, we identify the highest-ranking features associated with each mechanism and offer illustrative examples.

Related work

Conceptualizations of stigma

Goffman [18] influentially defined stigma as "an attribute that is deeply discrediting", and which reduces the stigmatized "from a whole and usual person to a tainted, discounted one" (p. 3). Goffman described stigma as a product of interactions, and stated that "a language of relations, not attributes, is really needed to describe stigma" [18] (p.3). The relational nature of stigma was emphasized by subsequent stigma theory [19, 20] that characterized stigma as a social process situated in a social context, with Link and Phelan [19] conceptualizing stigma as a convergence of labeling, stereotyping, separation, status loss, and discrimination, all within a power structure.

To complement existing societal-level conceptualizations of stigma with individual-level ones and create a more comprehensive theory of stigma and its impact, Earnshaw and Chaudoir [13] proposed the Stigma Framework. In this framework, which draws on stigma theory from a variety of domains [19–23], attention is given to both the mechanisms of stigma employed by those with power, and also the ways that stigma is experienced or adopted by stigmatized individuals. Earnshaw and Chaudoir distinguish three mechanisms employed by those who distance themselves from the "mark" of stigma: prejudice, stereotyping, and discrimination; and three mechanisms (hereafter primarily called "types") for those who experience stigma: *Internalized Stigma*, *Anticipated Stigma*, and *Enacted Stigma*. Table 1 provides definitions and examples of each of the three types of experienced stigma, in the context of substance use, as defined in Smith et al. [11]. The stigma mechanisms identified by the Stigma Framework have been assessed in various health-related contexts and have been associated with physical, mental, and behavioral outcomes for those that experience stigma [11, 24, 25].

Despite the existence of different conceptualizations of stigma, there is much that we do not yet understand

Table 1 Substance use stigma type definitions adapted from Smith et al. [11]

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."

about stigma processes. In particular, there is a recognized need to more clearly define and characterize the nature of stigma [9, 26]; to identify societal and individual-level factors affecting stereotyping, prejudice, and discrimination [12]; and to develop a more nuanced understanding of how different stigma mechanisms may affect substance use recovery [11]. In this study, we develop models to identify stigma in a large social media dataset for subsequent qualitative analysis intended to enhance our understanding of the complex interplay of the effects of stigma on the individual within their embedded contexts.

Computational models of stigma detection

Although a multitude of computational models for the detection of abusive language and hate speech in social media texts has been proposed [27, 28], the computational detection of social stigma has been less extensively explored. Whereas hate speech is commonly defined as a communicative act of disparagement of a person or group [29], 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 [1, 18, 19, 30]. Research on stigma detection in a variety of specific domains has been conducted, with works on the detection of depression stigma [14], mental health stigma [31, 32], stigmatizing language in healthcare discussions [33], Alzheimer's Disease stigma [34], schizophrenia stigma [35], and obesity stigma [36].

Li et al. [14] produce models 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 in conjunction with a simplified Chinese version of Linguistic Inquiry and Word Count (LIWC) features [37]. 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. [33] build 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.

Gottipati et al. [32] perform mental disorder stigma detection on a corpus of mental health-related news articles published by Singapore's largest media organizations. The authors create an (approximately) balanced dataset of stigmatizing and non-stigmatizing news article titles paired with a sentence from the same article. The researchers create features from TF-IDF weighted n-grams and compare a variety of machine learning classifiers, finding best performance with XGBoost [38].

To develop a model for detecting stigmatizing language related to mental health, Lee and Kyung [31] 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 [39] 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. We summarize the results of the four stigma detection studies described here in Table 2.

Although research on health-related stigma detection has been performed in a variety of domains, to our knowledge, all have treated stigma as a single monolithic concept. In this work, we incorporate the three stigma mechanisms (Internalized, Anticipated, and Enacted Stigma) of the Stigma Framework [13] to better differentiate between different types of stigma

Table 2 Summary of stigma detection studies

Authors	Features	Classifier	Stigma variant	Dataset type	Performance
Li et al. [14]	LIWC (Simplified Mandarin)	Random forest	Depression stigma	Imbalanced	0.752 F1
Straton et al. [33]	TF-IDF weighted n-grams + LIWC	CNN	Pro-vaccination stigma	Balanced	0.885 Accuracy
			Anti-vaccination stigma	Balanced	0.889 Accuracy
Gottipati et al. [32]	TF-IDF weighted n-grams	XGBoost	Mental disorder stigma	Balanced	0.772 Accuracy
Lee and Kyung [31]	BERT encodings	Single-layer neural network	Mental health stigma	Balanced, Synthetic	0.98 F1

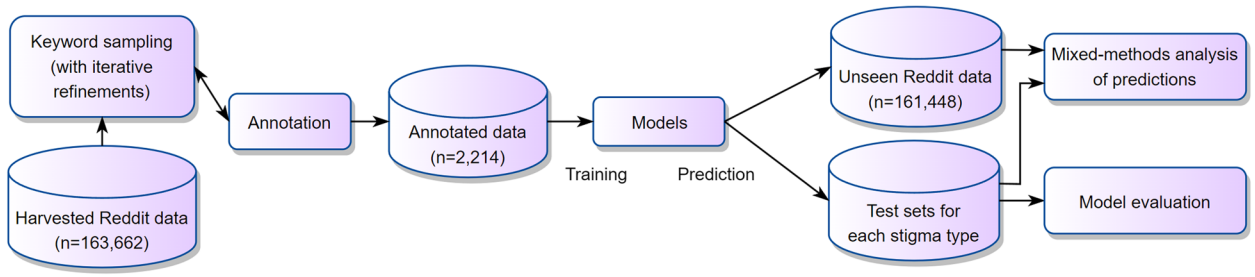

Fig. 1 Project overview flowchart

experiences, including identifying linguistic features which are most characteristic of each stigma type. For instance, the social media examples that we observed included stigmatizing language ("my sister is a hopeless alcoholic"), reports of stigmatization ("my husband took away the kids and said I'd never get clean"), and the experience of stigma ("I feel so much shame that I can't tell anyone").

Based on the effectiveness of BERT contextual embeddings, TF-IDF-weighted n-grams, and LIWC features for the purpose of stigmatizing language detection [14, 31, 33], we experiment with combinations of these resources. Given the prevalence of affect types such as sadness, anxiety, and fear in social media posts discussing experiences of substance use [40] and prior literature arguing that emotion regulation can be a factor in stigma coping [41, 42], we also experiment with count-based features derived from extant affect lexicons and our own handcrafted stigma lexicons. These handcrafted lexicons incorporate affective, social, and behavioral concepts based on stigma theory, including anxiety, depression, and secretive behavior [5, 9].

Methods

In this study, we employ classifiers to identify three different types of stigma in a social media dataset. We train and evaluate a set of models for each stigma type and then perform a mixed-methods analysis of the data identified by these models. A flowchart overview of our project is depicted in Fig. 1.

Dataset creation

Harvesting data

To create our dataset, approximately 160 thousand English-language Reddit posts authored between January 1, 2013 and December 31, 2019 were collected using Pushshift.io [43]. To capture diverse manifestations of substance use stigma and stigma-related behaviors (including navigation of legality for users), we focused on three substances for this analysis: alcohol, cannabis, and opioids. We selected subreddits related to the three substances of interest (e.g., 'r/stopdrinking', 'r/marijuana', and 'r/opiates') and sampled only thread-initiating posts, as these posts often contain richer descriptions of Redditor's experiences [44]. In our previous research [40, 45], we found these subreddits contained detailed accounts of both substance use and SUD recovery. Table 3 provides a breakdown of post counts for each subreddit in the harvested Reddit data. Subreddits that allude to or mention recovery or support in subreddit titles, descriptions or rules are labeled with checkmarks.

Sampling for annotation

We observed that posts containing explicit references to stigma were relatively uncommon. To increase the volume of relevant data for annotation and to support subsequent natural language processing, we employed the keyword sampling method used in Chen et al. [40] to build our annotated corpus. Only the posts that matched a regular expression containing a keyword list were sampled to increase the probability of sampling

Table 3 Subreddit distribution in harvested data

Substance focus	Subreddit	Support focus	Post count
Alcohol	r/alcohol		248
	r/cripplingalcoholism	✓	3,091
	r/stopdrinking	✓	24,879
Alcohol total			28,218
Cannabis	r/Marijuana		50,075
	r/Petioles		322
	r/cannabis		4,574
	r/leaves	✓	28,943
	r/trees		33,756
Cannabis total			117,670
Opioids	r/OpiatesRecovery	✓	7,489
	r/opiates		10,285
Opioids total			17,774
Total posts			163,662

Table 4 Keywords used in enrichment sampling

Category	Keyword list
Actor keywords	'family', 'friend(s)', 'parent(s)', 'everyone', 'co-worker(s)', 'coworker(s)', 'wife', 'husband', 'children', 'kid(s)', 'brother(s)', 'father', 'mother', 'dad', 'mom', 'girlfriend', 'boyfriend', 'bf', 'gf', 'doctor(s)', 'daughter(s)'
Stigma keywords	'rock bottom', 'dangerous', 'unpredictable', 'embarrass', 'shame', 'bias', 'prejudice','disappoint', 'weak', 'lazy', 'inadequate', 'untrustworthy', 'blame', 'hopeless', "stereotype", 'judg', 'discrimin', 'worthless', 'loser', 'failure', 'disgust', 'self-loathing', 'unclean', 'no future', 'trust', 'annoy', 'secret', 'hid'

stigma-related content. The theory-informed keyword list, derived from stigma literature [10, 11, 24, 25], 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'). Over the course of the annotation process, this list of keywords was iteratively refined to increase the prevalence of stigma in samples. The final set of sampling keywords is listed in Table 4. Additionally, subreddits that produced low yields for stigma content (e.g., r/alcohol, r/Petioles, r/trees) were removed from the candidates for annotation sampling. Table 5 shows the breakdown of post counts for each of the subreddits and the distribution of the three stigma types in the annotated dataset.

Annotation process

Three annotators with expertise in informatics, natural language processing, nursing, and public health annotated a total of 2,214 Reddit posts at the span-level for three stigma types based on the Stigma Framework [13]: 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 [11, 46]. A detailed description of our annotation guidelines is provided as Additional file 1.

Annotators independently identified passages containing stigma in the posts before discussing and reconciling the annotations. In addition to labeling stigma spans, annotators also labeled posts for substance type and the author's recovery outlook (positive, neutral, or negative), and identified spans containing mentions of

Table 5 Subreddit and stigma type distribution in annotated data ($n = 2,214$). Each post can contain multiple stigma types, and thus the sum of columns 4 through 6 can exceed the total post count in column 7

Substance focus	Subreddit	Support focus	Internalized Stigma	Anticipated Stigma	Enacted Stigma	Post count
Alcohol	r/cripplingalcoholism	✓	5 (62.50%)	1 (12.50%)	4 (50.00%)	8
	r/stopdrinking	✓	321 (47.77%)	132 (19.64%)	135 (20.09%)	672
Alcohol total			326 (47.94%)	133 (19.56%)	139 (20.44%)	680
Cannabis	r/Petioles		0 (00.00%)	0 (00.00%)	0 (00.00%)	5
	r/leaves	✓	208 (38.24%)	68 (12.50%)	51 (09.38%)	544
	r/trees		2 (07.14%)	1 (03.57%)	1 (03.57%)	28
Cannabis total			210 (36.40%)	69 (11.96%)	52 (09.01%)	577
Opioids	r/OpiatesRecovery	✓	204 (30.31%)	182 (27.04%)	126 (18.72%)	673
	r/opiates		24 (08.30%)	36 (12.46%)	44 (15.22%)	289
Opioids total			228 (23.70%)	218 (22.66%)	170 (17.67%)	962
Total posts			764 (34.51%)	420 (18.97%)	361(16.31%)	2,214

Table 6 Cohen's kappa scores for pairwise inter-annotator agreement (IAA) prior to adjudication

Pairings	Internal	Anticip	Enacted	Overall[a]
A1, A2	0.60	0.53	0.46	0.69
A1, A3	0.61	0.62	0.48	0.71
A2, A3	0.58	0.54	0.43	0.66
mean	0.60	0.56	0.46	0.69

[a] Including additional labels, such as substance type

social isolation and labels (e.g., 'addict'). Table 6 lists pairwise inter-annotator agreement for the three annotators at post level, prior to reconciliation, measured using Cohen's Kappa [47]. Overall, pair-wise agreement on the stigma mechanisms reflected moderate agreement [48], with the highest agreement being for Internalized Stigma. Pair-wise agreement scores on all annotation types varied between 0.66 and 0.71, indicating substantial agreement.

Text segmentation

In the annotated corpus, we observed that Reddit posts ranged in length from 28 characters to 25,743 characters, with a mean length of 1,816 characters (Fig. 2). As many posts exceed the 512-token input length limit of the RoBERTa encoder [49] that we use in our detection model, we opt to chunk posts into text segments. We use the term 'segment' to refer to the chunks of text used as inputs to our classifiers, and we use 'span' to refer to passages of text within posts labeled by annotators. We map the annotated span labels onto the segments, and then use the labeled 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 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 (Table 7). When splitting posts down to the level of sentences, we see severe class imbalance, with only 1.69% of the data containing Enacted Stigma.

Class imbalance can result in classifiers which perform well for the majority class, but poorly for the minority class [50, 51]. To mitigate class imbalance, we experimented with a variety of segmentation lengths, and found the best performing length to be approximately 600 characters. At this length, text segments seem to be short enough to mitigate the amount of irrelevant

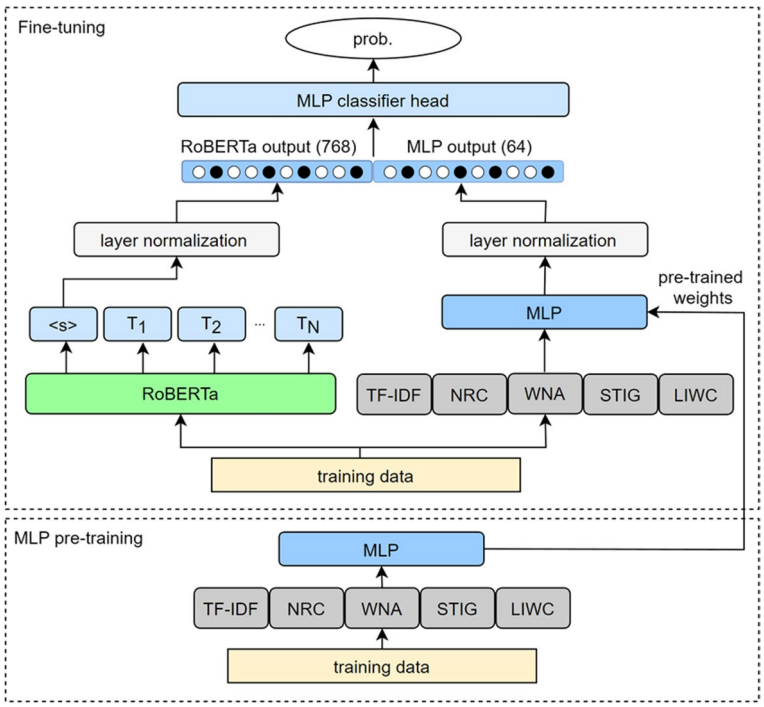

Fig. 2 Architecture of the hybrid model

Table 7 Stigma-positive portion of annotated corpus

Text level	Internalized Stigma n / %	Anticipated Stigma n / %	Enacted Stigma n / %	Total texts
Post	764 (34.51%)	420 (18.97%)	361 (16.31%)	2,214
Segment[a]	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

[a] Segments are ~ 600 characters in length

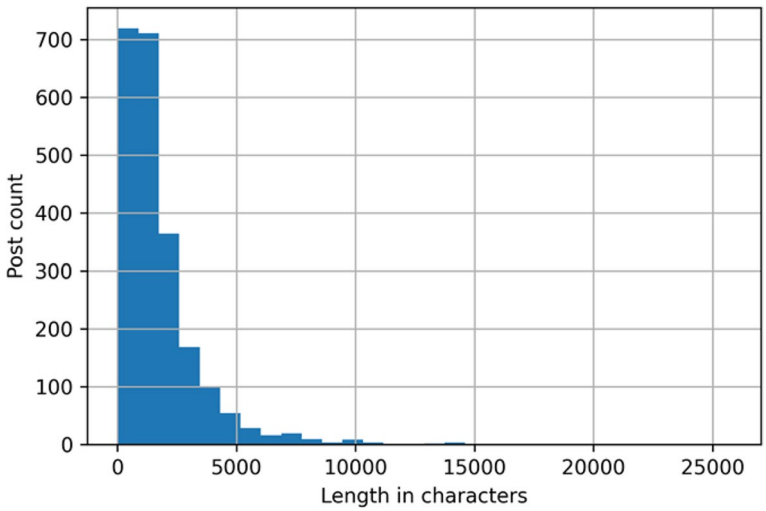

Fig. 3 Histogram of post character length

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 Natural Language Toolkit (NLTK) 3.5 [52]. 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.

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. Because each segment of input text may be stigma-positive for multiple stigma types, we treat this classification task as a set of independent binary classification tasks rather than a single multi-class classification task.

We utilize a RoBERTa encoder [49] as the main component of the classifier, and also make use of n-gram features, features derived from affective and psychological lexicons, and handcrafted features to enrich the model with external knowledge relevant to the task. To integrate RoBERTa embeddings with the additional features, we use a hybrid model (Fig. 3) based on Prakash et al. [53], 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[1] Emotional Intensity Lexicon [54], features derived from Wordnet-Affect [55], features generated from the LIWC 2015 lexicon [37], and handcrafted substance use stigma features.

After pre-training is complete, the trained MLP weights are used along with a pre-trained RoBERTa encoder in the fine-tuning process. The < s > token output of the RoBERTa encoder and the MLP output are normalized and then concatenated before being passed to an

[1] National Research Council Canada.

Table 8 Categories and concepts included in feature sets

Feature set	Categories and concepts
NRC Affective Intensity Lexicon	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
LIWC 2015	analytic, clout, authentic, tone, WPS, sixltr, dic, function, pronoun, ppron, i, we, you, shehe[a]
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

[a] See Pennebaker et al. (2015) [37] for full LIWC category list

MLP classifier head, which outputs the probability that a given sequence of text contains the current type of substance use stigma.

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 [56] to create TF-IDF weighted n-grams in the range (2, 6) with a dimensionality of 10,000.

NRC affective intensity features (NRC)

We include NRC features [54] 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 8). To produce feature vectors, we follow the method of Babanejad et al. [57], who create 'EAISe' representations (Emotion Affective Intensity with Sentiment Features) for their sarcasm detection model.

Wordnet Affect features (WNA)

Wordnet-Affect [1], developed based on Wordnet 1.6 [58], 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 8) and constructed 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.

LIWC features

Linguistic, grammatical, and psychological features are generated using LIWC 2015 software [37]. We remove the 'word count' feature and retain all others, resulting in a 92-dimensional vector.

Substance use stigma features (INT / ANT / ENA)

We create handcrafted lexicons (identified as 'INT', 'ANT', and 'ENA') to capture affective, behavioral, and social concepts related to each stigma type. These lexicons were developed through examination of TF-IDF weighted n-gram chi-square rankings for the training data, identification of recurring concepts in the stigma-positive examples of the training data that corresponded to concepts from stigma literature and survey instruments [10, 11, 24, 25, 46, 59], and iterative building and evaluation of lexical sets for each concept using a validation set. For Anticipated Stigma, an associated behavior such as concealment [25] 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 8, and the complete list of keywords included in each concept is listed in Additional file 2. 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 associated with the word.

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

Training

Data handling

Training sets are sampled from our segment-level data and contain a mixture of stigma-positive and stigma-negative texts. In development, the 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.

Hyperparameters

We train all models on a single Tesla A100 GPU on the Google Colab platform. Training is implemented using Pytorch 1.12 [60] and the Huggingface library [61]. 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. 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 min of GPU time were required to train a single hybrid model.

Evaluation

Model evaluation

As we sought to identify the stigma-positive Reddit posts within the unseen harvested Reddit data, we evaluate each model's predictions at the post-level by mapping segment predictions to each post. We compare the performance of models by reporting the mean macro F1 score of five runs on the same data, using different random seeds. We list results from variations of hybrid models utilizing different sets of features. As a baseline for comparison to the 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.

Improvements over the RoBERTa-only baselines are considered significant at a significance level (α) of 0.05 according to McNemar's test [62] with false discovery rate (FDR) correction [63]. McNemar's significance test has been considered appropriate for binary classification tasks [64]; thus, we employ it on the predictions of the paired models. Because we make multiple hypothesis tests in our comparisons, FDR correction is applied to p-values.

To explore each feature set's potential for use in stigma detection, we also considered the results of MLP evaluation on single feature sets and set combinations. We use an MLP for this comparison rather than a hybrid model since in the hybrid models, redundancies in the information encoded by feature set combinations and the information encoded by RoBERTa can make the relative performance contribution of each feature set difficult to disentangle. We also perform exploratory feature ranking of all features using the chi-square measure to explore the strength of association between each feature and its relevant stigma type. The feature selection tools of the Scikit-learn package were used to implement this experiment [56].

Last, we perform an error analysis of the hybrid model's predictions. This evaluation not only informs future improvements on our approach, but also provides insights into difficulties that arise in the perception and experience of stigma.

Mixed-methods analysis

Mixed-methods research can facilitate research that cannot be answered using a single method. Though there is controversy concerning what constitutes mixed-methods research, integrating quantitative and qualitative approaches is considered increasingly important, and extant literature has observed and demonstrated that the definition of mixed-methods research will continue to grow [65, 66]. In this study, we leverage both quantitative and qualitative methods for various affordances identified by Doyle et al. [66] including: triangulation, completeness, and illustration of data.

We performed a mixed-methods analysis to: 1) estimate the amount of stigma in the larger social media data store; and 2) characterize the nature of the different stigma mechanisms. First, we characterized the presence of stigma in the unseen portion of the harvested Reddit data by examining patterns in the distribution of stigma predictions with respect to substance and subreddit, and the correlations between stigma type predictions. We employ chi-square tests to compare the presence of the stigma mechanisms in the three substances. As a chi-square test of independence on its own merely shows that there is an association between two nominal variables and does not show which cells are contributing to the lack of fit [67, 68], we calculated standardized Pearson residuals. A standardized Pearson residual exceeding two in absolute value in a given cell indicates a lack of fit [67, 68]. Second, we considered the feature rankings and the instances of predicted stigma in the test data in concert to illustrate how the three types of stigma concretely manifest in cognitive and emotional processes, social interactions, and behaviors in everyday life. To protect the identities

Table 9 Post-level results across models and stigma types. Scores are macro F1 mean values of 5 runs (± std. dev.). 'STIG' refers to handcrafted stigma features, which are specific to each stigma type. [a]Indicates significant improvement over in-class baseline

Model	Features	Internalized	Anticipated	Enacted
RoBERTa-base	-	81.45 ± 2.36	74.34 ± 2.61	69.96 ± 3.82
MLP + RoBERTa-base	STIG	82.25 ± 1.63	75.22 ± 2.94	74.38[a] ± 1.37
	TF-IDF + NRC + WNA + LIWC	81.00 ± 1.16	75.68 ± 0.85	73.35[a] ± 2.54
	TF-IDF + NRC + WNA + LIWC + STIG	82.50 ± 2.28	76.30 ± 2.39	**75.62**[a] ± 1.28
RoBERTa-large	-	83.78 ± 1.16	71.60 ± 3.11	70.46 ± 2.64
MLP + RoBERTa-large	STIG	85.83[a] ± 1.17	**78.68**[a] ± 1.93	72.23 ± 3.13
	TF-IDF + NRC + WNA + LIWC	85.27 ± 1.56	75.55[a] ± 1.10	69.26 ± 1.67
	TF-IDF + NRC + WNA + LIWC + STIG	**86.68**[a] ± 0.68	77.03[a] ± 1.28	70.21 ± 2.05

Table 10 Post-level MLP results across features and stigma types. Scores are macro F1 mean values of 5 runs (± std. dev.). 'STIG' refers to handcrafted stigma features, which are specific to each stigma type. Bold values indicate the best result for each stigma type

Model	Features	Internalized	Anticipated	Enacted
MLP	TF-IDF	65.35 ± 0.54	57.94 ± 1.43	37.96 ± 6.01
	NRC	56.95 ± 3.72	35.43 ± 8.07	46.07 ± 6.26
	WNA	71.88 ± 2.46	42.00 ± 5.85	33.29 ± 11.41
	LIWC	58.20 ± 2.98	40.27 ± 1.48	60.24 ± 4.65
	STIG	**76.52** ± 0.74	**65.28** ± 1.65	57.09 ± 1.86
	TF-IDF + NRC	66.44 ± 0.83	57.84 ± 0.70	45.02 ± 3.45
	TF-IDF + NRC + WNA	71.8 ± 1.46	52.24 ± 1.45	49.14 ± 6.28
	TF-IDF + NRC + WNA + LIWC	73.34 ± 2.92	59.22 ± 1.16	60.43 ± 1.18
	TF-IDF + NRC + WNA + LIWC + STIG	73.58 ± 0.43	65.14 ± 2.02	**61.73** ± 1.42

of the posters, we employ synthetic quotes in our illustration [69].

Results and discussion

Model performance and evaluation

Overall model performance

Table 9 lists the results of post-level stigma detection for the three stigma types. For all three stigma types, we found hybrid models that significantly outperformed their respective RoBERTa-only baselines, with the largest gain observed for the Anticipated Stigma RoBERTa-large hybrid model using only the handcrafted stigma features (+7.08 F1). These results provide evidence that n-gram, affective, behavioral, and social features can be combined with contextual embeddings to improve substance use stigma detection.

In the results of MLP evaluation (Table 10), the handcrafted lexicons (STIG) appeared to be relatively effective resources for the task of stigma detection, and the other feature sets (NRC, WNA, and LIWC) also appear to be viable resources (to varying degrees). For individual feature sets, the handcrafted stigma lexicons appeared to provide the best results for Internalized Stigma and Anticipated Stigma, whereas LIWC provided best results for Enacted Stigma. For feature set combinations, adding additional feature sets usually led to improvement for MLP models (with some exceptions), although the combination of all features only outperformed the handcrafted stigma lexicons for the case of Enacted Stigma.

Comparing performance by stigma mechanisms and contributing features

The results in Tables 9 and 10 show that, 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 of all features (Table 11), count-based features had stronger associations (higher chi-square scores) with Internalized Stigma than they did with the other stigma types. Affective concepts such as 'shame' and 'guilt' had strong relationships with Internalized Stigma, which likely benefitted performance.

Overall performance for Anticipated and Enacted Stigma was weaker than for Internalized Stigma. There

Table 11 Top 15 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. *P*-values are listed with FDR (Benjamini-Hochberg) correction

	Internalized Stigma			Anticipated Stigma			Enacted Stigma		
Rank	**Feature**	**$\chi 2$**	***p***	**Feature**	**$\chi 2$**	***p***	**Feature**	**$\chi 2$**	***p***
1	INT_shame	544.863	1.7e-116	ANT_secrecy	148.754	3.2e-30	ENA_labels	54.512	1.6e-9
2	WNA_guilt	218.519	9.6e-46	ANT_social	29.979	2.2e-4	ENA_stigmatizing_actions	42.759	3.1e-7
3	WNA_shame	169.611	3.0e-35	i lied	22.762	0.006	ENA_trust	24.820	0.002
4	INT_self_blame	112.78	6.1e-23	i hid	21.309	0.008	LIWC_Tone	22.350	0.006
5	LIWC_Tone	99.605	3.8e-20	LIWC_Tone	21.205	0.008	ENA_loss	17.351	0.063
6	INT_despair	55.899	1.3e-10	ANT_status	20.109	0.012	my sister	14.456	0.242
7	INT_pejoratives	53.683	3.4e-10	i hiding	19.100	0.016	recovering alcoholic	13.718	0.307
8	INT_labels	51.129	1.1e-9	secret i	19.016	0.016	give shit	11.979	0.681
9	NRC_negative	49.785	1.9e-9	i hide	18.245	0.022	low key	11.607	0.739
10	LIWC_Clout	47.172	6.6e-9	hide family	16.083	0.061	my husband	11.038	0.904
11	INT_loss	41.772	9.4e-8	i hidden	15.746	0.062	empty bottles	10.218	0.921
12	shame guilt	35.217	2.5e-6	track marks	15.710	0.062	treated like	10.068	0.921
13	i ashamed	33.901	4.5e-6	knows i	14.124	0.124	drug addicts	9.618	0.921
14	ashamed i	26.954	1.5e-4	i tell anyone	14.113	0.124	ENA_punishment	9.023	0.921
15	the shame	25.334	3.3e-4	one knows	13.720	0.143	think going	8.617	0.921
	NRC Affective Intensity Lexicon feature								
	Wordnet-Affect feature								
	Substance use stigma feature (INT/ANT/ENA)								
	LIWC 2015 feature								

may be a number of reasons for this. First, Anticipated and Enacted Stigma had fewer examples and relatively weaker associations with count-based features in comparison with Internalized Stigma. For Enacted Stigma, the highest-ranking features were labels such as 'alcoholic' and 'junkie', which were fairly common in the entire corpus. Labeling terms such as 'alcoholic' may be used to enact stigma, but they may also be used to express membership in recovery groups and are a part of 'recovery dialects' used within such groups [2]. Moreover, labeling terms may also be appropriated by members of stigmatized groups to increase perceptions of power for the stigmatized individual or group [70]. The variety of motivations behind the uses of such labeling terms such as 'junkie' may be a limiting factor to their viability as features for stigma detection.

Another potential factor for the weaker performance for Anticipated and Enacted Stigma may be their social nature. Whereas Internalized Stigma frequently focus on a single entity (the post author), with feature rankings showing strong relationships with inward features (n-grams such as 'i ashamed'), both Anticipated and Enacted Stigma involved other actors. With Anticipated Stigma, the highest ranking features involved concealment of use (ANT_secrecy) and other actors (ANT_social), as post authors were concerned about concealing their use from others. With Enacted Stigma, there was a wide variety of actors involved in the relationships between the stigmatizer and the person(s) being stigmatized (e.g. 'family to partner', 'partner to society', 'coworkers to society'). Further, while Internalized Stigma frequently focused on the act of shaming oneself, Enacted Stigma involved a more diverse set of verbs/actions through which stigma was performed (e.g., disapproving looks, expressions of distrust, arrests, searches, evictions, insults, generalizations, coerced drug tests, denial of healthcare services, termination of employment, termination of personal relationships). Many of the verbs related to these stigmatizing actions were included in the ENA_stigmatizing_actions and ENA_trust features, which ranked second and third, respectively, in the feature ranking.

Model performance by stigma type followed a similar pattern to that of inter-annotator agreement across stigma types (Table 6), in which annotators found highest agreement on Internalized Stigma and less agreement on Anticipated and Enacted Stigma. The complexities involved in identifying these two stigma types seemed to be a challenge for both human annotators as well as the models.

Error analysis

We provide an error analysis of the Anticipated and Enacted Stigma models to gain insights into the challenges involved in detecting these stigma types. We give synthetic quotes based on our data to demonstrate error types, with features typical of Anticipate or Enacted Stigma texts bolded.

Temporal errors We observed that both the Anticipated and Enacted Stigma hybrid models produced false positives for texts which do not match the temporal

Table 12 Subreddit and predicted stigma type distribution in the unseen data ($n = 161,448$). Each post can contain multiple stigma types, and thus the sum of columns 4 through 6 can exceed the total post count in column 7

Substance focus type	Subreddit	Support focus	Internalized Stigma	Anticipated Stigma	Enacted Stigma	Post count
Alcohol	r/alcohol		2 (0.81%)	3 (1.21%)	3 (1.21%)	248
	r/cripplingalcoholism	✓	160 (5.19%)	87 (2.82%)	251 (8.14%)	3,083
	r/stopdrinking	✓	2,342 (9.67%)	1,119 (4.62%)	1,702 (7.03%)	24,207
Alcohol total			2,504 (9.09%)	1,209 (4.39%)	1,956 (7.10%)	27,538
Cannabis	r/Marijuana		41 (0.08%)	68 (0.14%)	374 (0.75%)	50,075
	r/Petioles		5 (1.58%)	3 (0.95%)	6 (1.89%)	317
	r/cannabis		0 (0.00%)	0 (0.00%)	9 (0.20%)	4,574
	r/leaves	✓	1,742 (6.13%)	686 (2.42%)	544 (1.92%)	28,399
	r/trees		45 (0.13%)	139 (0.41%)	210 (0.62%)	33,733
Cannabis total			1,833 (1.57%)	896 (0.77%)	1,143 (0.98%)	117,098
Opioids	r/OpiatesRecovery	✓	359 (5.27%)	407 (5.97%)	250 (3.67%)	6,816
	r/opiates		83 (0.83%)	98 (0.98%)	134 (1.34%)	9,996
Opioids total			442 (2.62%)	505 (3.00%)	384 (2.28%)	16,812
All posts			4,779 (2.96%)	2,610 (1.62%)	3,483 (2.16%)	161,448

requirements of their respective stigma type (future for Anticipated Stigma, present or past for Enacted Stigma). The following example (a false positive for Enacted Stigma due to temporal mismatch), is representative of this error type:

> *If I come clean, my family will disown me – that isn't even an option.*

For the RoBERTa-only baseline models, 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 [71].

Stigmatizing quitters During annotation, we observed that individuals abstaining from substance use were pressured by persons who engaged in substance use, often in the context of alcohol use when it is normalized in home or work-related settings. Though this behavior was not annotated as 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. 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 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.

Motivations Both the baseline and hybrid models for Anticipated and Enacted Stigma were prone to produce false positives for texts where typical features of stigma are present, but the motivation behind an action potentially construed as stigmatizing is unrelated to stereotyping, prejudice, or discrimination. In the following example, a partner appears to terminate a relationship due to apathetic behavior rather than stigma, and thus should be labeled as stigma-negative:

> *I struggled for a long time with the sadness that comes with addiction, so the feelings of apathy that followed it seemed like a relief. Eventually, they resulted in my partner breaking up with me.*

Although BERT models have been demonstrated to encode information that can be leveraged to make predictions about causality [72], 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.

Characterizing the presence of stigma in the unexplored data

To better understand how the three stigma mechanisms outlined in the Stigma Framework manifest within our social media dataset, we employed the classifiers to identify instances of the stigma types in the previously unexplored portion of our collected Reddit data ($n = 161,448$). The distribution of stigma predictions across subreddits is presented in Table 12. Overall, the portion of stigma-positive predictions for each type were noticeably lower than the portions seen in the annotated data (Table 5).

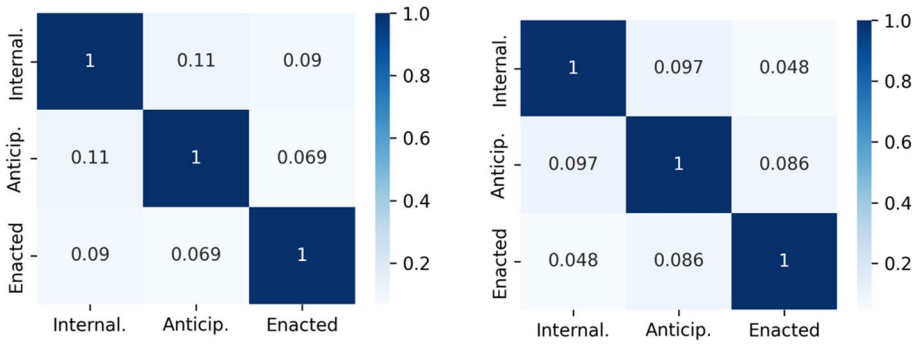

Fig. 4 Pearson correlation between stigma types for text segments in the annotated dataset (left) and the predictions on the unseen data (right)

This outcome aligns with expectations, given that: 1) keyword sampling was used to increase the proportion of stigma in the annotated data; and 2) in the unexplored data, a larger portion of posts originated from subreddits focused on general substance use, rather than on support or recovery. In both the predictions and annotations, we observed that, for all three substance types, the estimated stigma proportion was highest for support-focused subreddits, where posters often described challenging experiences relating to their attempts at recovery.

With respect to alcohol and cannabis, Internalized Stigma appeared to be the most common of the three stigma types. The focus on the self makes intuitive sense given the first-person viewpoint of social media narratives, and the prominent features of Internalized Stigma (Table 11) suggest that these data could serve as a rich source for future research on how individuals may seek to internally reconcile the cognitive and emotional aspects of shame and guilt that accompany Internalized Stigma.

However, in the case of opioids, we observed a higher frequency of Anticipated Stigma compared to Internalized Stigma. Chi-square tests examining the presence of the three stigma mechanisms in the three substances, with the standardized Pearson residual for Anticipated Stigma x Opioids, also confirm that the observed presence of Anticipated Stigma exceeds the expected in that case (see Additional file 3).

Co-occurrence of the three stigma mechanisms

We also explored the extent to which the stigma mechanisms co-occurred in the data. Figure 4 shows a Pearson correlation matrix between stigma labels for text segments in the annotated data (left) and also for the predictions on the unseen data (right). The largest correlation score is a value of 0.11 between Internalized and Anticipated Stigma (in the annotated data), indicating that text segments with multiple stigma labels are relatively infrequent in the annotated data. Although we observed some concepts were shared across stigma types in the feature rankings, such as labeling terms (e.g., 'addict'), the relatively low correlation between paired stigma types illustrates the utility of developing separate models for each stigma type. Furthermore, this underscores the potential utility of the three stigma types distinguished by the Stigma Framework [13] for future research in clarifying the mechanisms by which stigma can affect persons with SUDs.

Exploring the relationship between language and stigma experience

To characterize the nature of stigma as manifested in social media, we consider the feature rankings associated with each stigma type, along with the instances of stigma in the test data. Figure 5 depicts the concepts from the handcrafted stigma lexicons that were among the highest-ranking features for each stigma type, along with synthetic examples. Among the posts associated with Internalized Stigma, we observed an abundance of affective content (shame, self-blame, and despair). Our examination of the test data further uncovers that posts containing shame and self-blame also often involved the poster using self-deprecating language (in the form of pejoratives) and labels to describe themselves, and express feelings of weakness and perceptions of failure.

For Anticipated and Enacted Stigma, emotion was still important, but social and behavioral features were also prominent (i.e., ANT_social, ENA_stigmatizing_actions). The 'ANT_social' lexicon includes possible members of a user's social circle (e.g., 'parents', 'partner', 'friend'). Since, by definition, Internalized Stigma is focused on the self, Anticipated Stigma is focused on one's expectation of how they are perceived by others, and Enacted stigma, by stigmatizing behavior, these associations make intuitive sense. The social media data highlights additional features tied to Anticipated Stigma, such as secretive behavior, concern

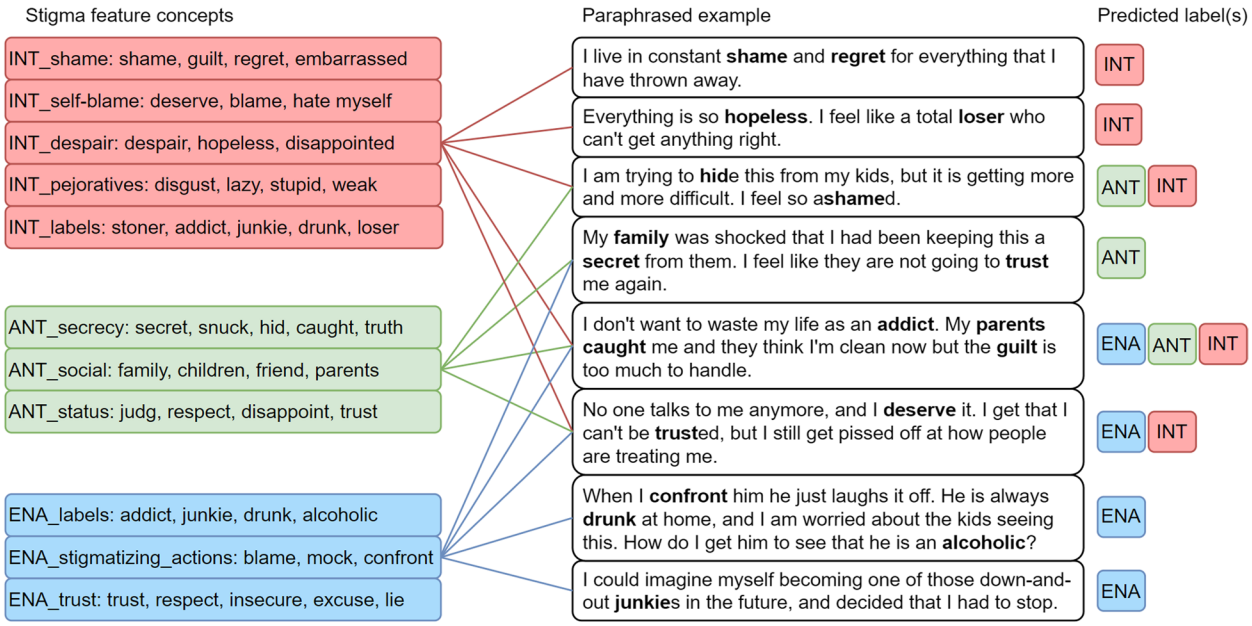

Fig. 5 Conceptual differentiation of stigma types. All examples are synthetic quotes that resemble the phenomena and sentiment observed in the data

over how one is perceived, and a fear of disappointing others. Notably, the theme of concealment, especially from close relations like family members, partners, or employers, is prominent in the Anticipated Stigma texts (as exemplified in the examples 3–5 in Fig. 5).

Enacted Stigma often involved the use of labels to describe another person, and as seen in the final two examples of Fig. 5, the usage of these terms can be descriptive ('He is always drunk') or may have judgmental motivations in their usage ('down-and-out junkies'). Stigmatizing actions related to judging, disparaging, or confronting others figured prominently in terms of this type of stigma, and could involve many different pairs of stigmatizer and stigmatized persons (e.g., parent–child, child-parent, friends, partners, co-workers, and the poster feeling stigmatized by the public, people, or society at large). Features related to trust also ranked highly for Enacted Stigma, corresponding to previous stigma research which identified 'untrustworthiness' as a common stereotype espoused by user's family members [24].

Other phenomena to consider were instances in which multiple stigma types were present. The third text in Fig. 5 exemplifies a common scenario for the pairing of Internalized Stigma and Anticipated Stigma, with posters expressing reticence to interact with others due to their own shame. Text segments containing all three stigma types were relatively rare in the annotated corpus (0.78% of all stigma-positive segments), though the fifth example in Fig. 5 illustrates an instance where an author appears to negatively judge persons experiencing SUDs, describe concealment of their own use, and express internal guilt for their use, all within a relatively brief sequence of text.

LIWC features

Similar to Straton et al. [33], 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. The clout feature, derived from ratios of personal pronoun frequencies, is based on Kacewicz et al. [73], who found that high-status authors consistently used more 1st person plural (e.g., 'we', 'our') and 2nd person singular ('you') pronouns, whereas low-status authors were more frequently self-focused and used more 1st person singular pronouns ('I', 'me'). This may explain the effectiveness of the clout feature for predicting Internalized Stigma (low clout scores appeared to be indicative of Internalized Stigma), which is heavily focused on inner experiences, with heavy use of 1st person pronouns. The LIWC emotional tone feature [74] calculates the difference between positive emotion word count and negative emotion word count, with higher scores indicating greater overall positivity. The generally negative emotional content of stigma-positive

texts is a likely factor for the high ranking of the tone feature for all three stigma types.

Discussion and limitations

In this study, our objective was to investigate how the three different stigma mechanisms in the Stigma Framework manifest differently in terms of distribution and nature in a social media dataset. Through an analysis of feature rankings, the distribution of predictions, and specific instances of stigma in our data, we discerned distinct patterns across Internalized, Anticipated, and Enacted Stigma. Furthermore, we characterized the language used to convey and describe each of these three mechanisms.

In terms of the distributions of the three stigma mechanisms, we observed that Internalized Stigma was the most prevalent stigma type with respect to alcohol and cannabis. However, in the case of opioids, Anticipated Stigma was more frequent than Internalized Stigma. Though these patterns were only observed in a single dataset and further exploration of the presence of different stigma mechanisms in other data is needed, it is worthwhile to consider these findings in the context of the larger societal concern about opioid use. Extant literature emphasizes that great care must be used in crafting public health messaging concerning opioid addiction due to the potential for increased stigmatization of those who use opioids [75]. The social environment surrounding opioid use appears to lead to greater anticipation of stigma and a tendency to conceal behavior, compared to the environments surrounding cannabis and alcohol. Thus, it may be important to focus on the portrayal of opioid use, anonymous forms of support, and an emphasis on support for interpersonal interactions in the context of opioid use.

Additionally, our study considered the nature of language used to express stigma as it manifests in social media. This exploration not only confirms that language is a powerful vehicle for expressing stigma, as established in prior literature [2], but also illuminates the nuanced relationship between word usage and specific stigma types, and the pivotal roles of affect, social perceptions, personal interactions, and behavior in the expression of stigma, in social media. In the social media data, we found that Internalized Stigma is predominantly characterized by emotional content, with a focus on shame, self-blame, and despair. In contrast, Enacted Stigma and Anticipated involve a complex interplay of emotional, social, and behavioral features. The former encompasses stigmatizing behaviors and issues of trust, while the latter centers on expectations of external perceptions and the fear of disappointing others. For Anticipated Stigma, the feature analysis demonstrated that issues of concealment were prominent, along with the presence of close interpersonal relationships.

Insights from this study can serve as priorities in the design of stigma reduction interventions. For example, the high-ranking features from the Enacted Stigma lexicon include both stigmatizing actions such as confronting and blaming, as well as indicators of trust (e.g., expressed as disappointment, suspicion, or a lack of respect for privacy). In future intervention development, the integration of components addressing these core issues is critical.

Overall, our findings improve our understanding of stigma mechanisms in social media discourse and could also inform the development of targeted interventions that address the challenges of those affected by stigma. Furthermore, the adaptability of our lexicons to stigma research in other contexts, such as HIV/AIDS or disordered eating, where similar emotions, behaviors (e.g., hiding, concealment), and attitudinal constructs such as trust [24, 76] are at play, hold promise for broader applications beyond substance use.

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 of its limitations. By sampling from a limited set of subreddits focused on substance use, 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. When making predictions on random samples, our models may have faced performance issues due to the increased imbalance between stigma-positive and stigma-negative texts.

To facilitate the aims of this research, we sought to identify stigma and accounts of stigma within social media narratives. In many of the possible instances of stigma that appear, 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. Collection and triangulation of data collected through other means, such as interview, survey, or diary data, could perhaps complement insights from social media.

Conclusion

In this study, we performed an examination of stigma surrounding substance use within the realm of social media. Our approach encompassed data collection, corpus annotation, and the development of binary classifiers tailored to detect three different stigma mechanisms. By synergizing contextual embeddings with count-based features, we achieved models that exhibited superior performance across all three stigma categories compared to RoBERTa-only baselines. Through a mixed-methods analysis of the model's predictions, we unraveled critical insights into the relations of word usage to the expression of different types of stigma. Affective, social, and behavioral features emerged as pivotal components in the expression of substance use stigma.

Our main contributions include: demonstrating a theory-based approach to extracting and comparing different types of stigma in a large social media dataset, and employing patterns in word usage to explore and characterize its manifestations. The insights from this study highlight the need to consider the impacts of stigma differently by mechanism (internalized, anticipated, and enacted), and enhance our current understandings of how the stigma mechanisms manifest within language in particular cognitive, emotional, social, and behavioral aspects. Moving forward, we envisage further analysis of stigma instances in our dataset to glean insights into how individuals navigate the challenges they encounter, informing the development of more effective stigma reduction strategies. Furthermore, the concepts encapsulated in our handcrafted lexicons hold promise for future stigma research in diverse contexts, extending the applicability of our findings beyond substance use disorders.

Abbreviations

SUD	Substance use disorder
MLP	Multi-layer perceptron
LIWC	Linguistic inquiry and word count
TF-IDF	Term frequency-inverse document frequency
BERT	Bidirectional encoder representations from transformers
RoBERTa	Robustly optimized bidirectional encoder representations from transformers
NRC	National research council Canada
NLTK	Natural language toolkit
WNA	Wordnet-affect
INT	Internalized stigma feature lexicon
ANT	Anticipated stigma feature lexicon
ENA	Enacted stigma feature lexicon

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s44247-024-00065-0.

Additional file 1. A detailed description of our annotation guidelines.

Additional file 2. A complete list of keywords included in each of the handcrafted stigma lexicons.

Additional file 3. Results of chi-square tests examining the distribution of stigma labels for each substance.

Acknowledgements
Not applicable.

Authors' contributions
ATC and DR conceptualized the study. All authors performed data curation, and DR and ATC performed data analysis. DR drafted the initial manuscript and iteratively revised with ATC. All authors reviewed and approved the final manuscript.

Funding
Research reported in this publication was supported by the National Institute On Drug Abuse of the National Institutes of Health under Award Number R21DA056684. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Availability of data and materials
Stigma datasets and models trained to detect stigma could potentially be used by bad actors to target vulnerable individuals. In order to reduce the risk of any potential harms to the authors of the sensitive posts examined in our research, we do not share our models or annotated dataset publicly.

Declarations

Ethics approval and consent to participate
The work performed in this study was determined as non-human subjects research by the Human Subjects Division at the University of Washington (STUDY00015737), and approved by the Office of Research Ethics and Integrity of the University of Melbourne (reference no. 2022–25512-34338–4). All methods were carried out in accordance with relevant guidelines and regulations.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]University of Washington School of Medicine, Biomedical Informatics and Medical Education, 850 Republican St., Box 358047, Seattle, WA 98109, USA. [2]School of Nursing, University of Washington, Box 357260, Seattle, WA 98195, USA. [3]School of Computing and Information Systems, University of Melbourne, Parkville, VIC 3052, Australia.

Received: 19 May 2023 Accepted: 26 January 2024
Published: 16 August 2024

References

1. Kulesza M, Ramsey S, Brown R, Larimer M. Stigma among Individuals with Substance Use Disorders: Does it Predict Substance Use, and Does it Diminish with Treatment? J Addict Behav Ther Rehabil. 2014;3(1):1000115. https://doi.org/10.4172/2324-9005.1000115.
2. Ashford RD, Brown AM, Ashford A, Curtis B. Recovery dialects: A pilot study of stigmatizing and nonstigmatizing label use by individuals in recovery from substance use disorders. Exp Clin Psychopharmacol. 2019;27(6):530–5. https://doi.org/10.1037/pha0000286.
3. Bozdağ N, Çuhadar D. Internalized stigma, self-efficacy and treatment motivation in patients with substance use disorders. J Subst Use. 2022;27(2):174–80. https://doi.org/10.1080/14659891.2021.1916846.
4. Hammarlund R, Crapanzano KA, Luce L, Mulligan L, Ward KM. Review of the effects of self-stigma and perceived social stigma on the treatment-seeking decisions of individuals with drug- and alcohol-use disorders. Subst Abuse Rehabil. 2018;9:115–36. https://doi.org/10.2147/SAR.S183256.
5. Livingston JD, Milne T, Fang ML, Amari E. The effectiveness of interventions for reducing stigma related to substance use disorders: a systematic review. Addiction. 2012;107(1):39–50. https://doi.org/10.1111/j.1360-0443.2011.03601.x.

6. Ashford RD, Brown AM, Curtis B. Substance use, recovery, and linguistics: The impact of word choice on explicit and implicit bias. Drug Alcohol Depend. 2018;189:131–8. https://doi.org/10.1080/07347324.2019.1585216.
7. Volkow ND, Gordon JA, Koob GF. Choosing appropriate language to reduce the stigma around mental illness and substance use disorders. Neuropsychopharmacol. 2021;46(13):2230–2. https://doi.org/10.1038/s41386-021-01069-4.
8. Brown SA. Standardized measures for substance use stigma. Drug Alcohol Depend. 2011;116(1):137–41. https://doi.org/10.1016/j.drugalcdep.2010.12.005.
9. Kulesza M, Larimer ME, Rao D. Substance Use Related Stigma: What we Know and the Way Forward. J Addict Behav Ther Rehabil. 2013;2(2). https://doi.org/10.4172/2324-9005.1000106.
10. Kulesza M, Watkins KE, Ober AJ, Osilla KC, Ewing B. Internalized stigma as an independent risk factor for substance use problems among primary care patients: Rationale and preliminary support. Drug Alcohol Depend. 2017;180:52–5. https://doi.org/10.1016/j.drugalcdep.2017.08.002.
11. Smith LR, Earnshaw VA, Copenhaver MM, Cunningham CO. Substance use stigma: Reliability and validity of a theory-based scale for substance-using populations. Drug Alcohol Depend. 2016;162:34–43. https://doi.org/10.1016/j.drugalcdep.2016.02.019.
12. Corrigan P, Schomerus G, Shuman V, Kraus D, Perlick D, Harnish A, et al. Developing a research agenda for understanding the stigma of addictions Part I: Lessons from the Mental Health Stigma Literature. Am J Addict. 2017;26(1):59–66. https://doi.org/10.1111/ajad.12458.
13. Earnshaw VA, Chaudoir SR. From Conceptualizing to Measuring HIV Stigma: A Review of HIV Stigma Mechanism Measures. AIDS Behav. 2009;13(6):1160–77. https://doi.org/10.1007/s10461-009-9593-3.
14. Li A, Jiao D, Zhu T. Detecting depression stigma on social media: A linguistic analysis. J Affect Disord. 2018;232:358–62. https://doi.org/10.1016/j.jad.2018.02.087.
15. Li A, Jiao D, Liu X, Zhu T. A Comparison of the Psycholinguistic Styles of Schizophrenia-Related Stigma and Depression-Related Stigma on Social Media: Content Analysis. J Med Internet Res. 2020;22(4): e16470. https://doi.org/10.2196/16470.
16. Clark O, Lee MM, Jingree ML, O'Dwyer E, Yue Y, Marrero A, et al. Weight Stigma and Social Media: Evidence and Public Health Solutions. Front Nutr. 2021;8:739056.
17. Dredze M. How Social Media Will Change Public Health. IEEE Intell Syst. 2012;27(4):81–4. https://doi.org/10.1109/MIS.2012.76.
18. Goffman E. Stigma: Notes on the management of spoiled identity. New York: Simon and Schuster; 2009. p. 52, 65.
19. Link BG, Phelan JC. Conceptualizing Stigma. Annu Rev Sociol. 2001;27(1):363–85. https://doi.org/10.1146/annurev.soc.27.1.363.
20. Parker R, Aggleton P. HIV and AIDS-related stigma and discrimination: a conceptual framework and implications for action. Soc Sci Med. 2003;57(1):13–24. https://doi.org/10.1016/S0277-9536(02)00304-0.
21. Brewer MB, Brown RJ. Intergroup relations. In: Gilbert DT, Fiske ST, Lindzey G, editors. The handbook of social psychology. 4. New York: Oxford University Press; 1998.
22. Meyer IH. Prejudice, Social Stress, and Mental Health in Lesbian, Gay, and Bisexual Populations: Conceptual Issues and Research Evidence. Psychol Bull. 2003;129(5):674–97. https://doi.org/10.1037/0033-2909.129.5.674.
23. Phelan J, Link BG, Dovidio JF. Stigma and Prejudice: One Animal or Two? Soc Sci Med. 2008;67(3):358–67. https://doi.org/10.1016/j.socscimed.2008.03.022.
24. Earnshaw V, Smith L, Copenhaver M. Drug Addiction Stigma in the Context of Methadone Maintenance Therapy: An Investigation into Understudied Sources of Stigma. Int J Ment Health Addict. 2013;11(1):110–22. https://doi.org/10.1007/s11469-012-9402-5.
25. Quinn DM, Earnshaw VA. Concealable Stigmatized Identities and Psychological Well-Being. Soc Personal Psychol Compass. 2013;7(1):40–51 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3664915/).
26. Crapanzano KA, Hammarlund R, Ahmad B, Hunsinger N, Kullar R. The association between perceived stigma and substance use disorder treatment outcomes: a review. Subst Abuse Rehabil. 2018;10:1–12. https://doi.org/10.2147/SAR.S183252.
27. Schmidt A, Wiegand M. A Survey on Hate Speech Detection using Natural Language Processing. In: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media. Valencia, Spain: Association for Computational Linguistics; 2017. p. 1–10. https://doi.org/10.18653/v1/W17-1101.
28. Yin W, Zubiaga A. Towards generalisable hate speech detection: a review on obstacles and solutions. PeerJ Comput Sci. 2021;7:e598. https://doi.org/10.7717/peerj-cs.598.
29. Nockleby JT, Levy LW, Karst KL, Mahoney DJ. Encyclopedia of the American constitution. Detroit, MI: Macmillan Reference; 2000. p. 1277–9.
30. Allport GW, Clark K, Pettigrew T. The Nature of Prejudice. 1954.
31. Lee MH, Kyung R. Mental Health Stigma and Natural Language Processing: Two Enigmas Through the Lens of a Limited Corpus. In: 2022 IEEE World AI IoT Congress (AIIoT). 2022. p. 688–91. https://doi.org/10.1109/AIIoT54504.2022.9817362.
32. Gottipati S, Chong M, Kiat A, Kawidiredjo B. Exploring Media Portrayals of People with Mental Disorders using NLP: In: Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies. p. 708–15. https://doi.org/10.5220/0010380007080715.
33. Straton N, Jang H, Ng R. Stigma Annotation Scheme and Stigmatized Language Detection in Health-Care Discussions on Social Media. In: Proceedings of the 12th Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association; 2020. p. 1178–90 https://aclanthology.org/2020.lrec-1.148.
34. Oscar N, Fox PA, Croucher R, Wernick R, Keune J, Hooker K. Machine Learning, Sentiment Analysis, and Tweets: An Examination of Alzheimer's Disease Stigma on Twitter. J Gerontol B. 2017;72(5):742–51. https://doi.org/10.1093/geronb/gbx014.
35. Jilka S, Odoi CM, van Bilsen J, Morris D, Erturk S, Cummins N, et al. Identifying schizophrenia stigma on Twitter: a proof of principle model using service user supervised machine learning. Schizophr. 2022;8(1):1–8. https://doi.org/10.1038/s41537-021-00197-6.
36. Pollack CC, Emond JA, O'Malley AJ, Byrd A, Green P, Miller KE, et al. Characterizing the Prevalence of Obesity Misinformation, Factual Content, Stigma, and Positivity on the Social Media Platform Reddit Between 2011 and 2019: Infodemiology Study. Journal of Medical Internet Research. 2022;24(12):e36729. Available from: https://www.jmir.org/2022/12/e36729.
37. Pennebaker JW, Boyd RL, Jordan K, Blackburn K. The Development and Psychometric Properties of LIWC2015. 2015. https://repositories.lib.utexas.edu/handle/2152/31333.
38. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. p. 785–94. Available from: http://arxiv.org/abs/1603.02754. Accessed 14 Jan 2022.
39. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:181004805. 2019. http://arxiv.org/abs/1810.04805. Accessed 11 May 2020.
40. Chen AT, Johnny S, Conway M. Examining stigma relating to substance use and contextual factors in social media discussions. Drug and Alcohol Dependence Reports. 2022;3:100061. https://doi.org/10.1016/j.dadr.2022.100061.
41. Hatzenbuehler ML, Nolen-Hoeksema S, Dovidio J. How Does Stigma "Get Under the Skin"?: The Mediating Role of Emotion Regulation. Psychol Sci. 2009;20(10):1282–9. https://doi.org/10.1111/j.1467-9280.2009.02441.x.
42. Wang K, Burton CL, Pachankis JE. Depression and Substance Use: Towards the Development of an Emotion Regulation Model of Stigma Coping. Subst Use Misuse. 2018;53(5):859–66. https://doi.org/10.1080/10826084.2017.1391011.
43. Baumgartner J, Zannettou S, Keegan B, Squire M, Blackburn J. The Pushshift Reddit Dataset. Proc Int AAAI Conf Web Soc Media. 2020;14:830–9.
44. MacLean D, Gupta S, Lembke A, Manning C, Heer J. Forum77: An Analysis of an Online Health Forum Dedicated to Addiction Recovery. In: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing. New York, NY, USA: Association for Computing Machinery; 2015. p. 1511–26. (CSCW '15). https://doi.org/10.1007/BF02295996.
45. Benson R, Hu M, Chen AT, Zhu SH, Conway M. Examining Cannabis, Tobacco, and Vaping Discourse on Reddit: An Exploratory Approach Using Natural Language Processing. Front Public Health. 2022. Available from: https://www.frontiersin.org/article/10.3389/fpubh.2021.738513.
46. Palamar JJ, Kiang MV, Halkitis PN. Development and Psychometric Evaluation of Scales that Assess Stigma Associated With Illicit Drug Users. Subst

Use Misuse. 2011;46(12):1457–67. https://doi.org/10.3109/10826084.2011.596606.
47. Cohen J. A Coefficient of Agreement for Nominal Scales. Educ Psychol Measur. 1960;20(1):37–46. https://doi.org/10.1177/001316446002000104.
48. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med. 2005;37(5):360–3.
49. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv. 2019. https://doi.org/10.48550/arXiv.1907.11692.
50. He H, Garcia EA. Learning from Imbalanced Data. IEEE Trans Knowl Data Eng. 2009;21(9):1263–84. https://doi.org/10.1109/TKDE.2008.239.
51. Haixiang G, Yijing L, Shang J, Mingyun G, Yuanyue H, Bing G. Learning from class-imbalanced data: Review of methods and applications. Expert Syst Appl. 2017;73:220–39. https://doi.org/10.1016/j.eswa.2016.12.035.
52. Bird S, Klein E, Loper E. Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc. 2009.
53. Prakash A, Tayyar Madabushi H. Incorporating Count-Based Features into Pre-Trained Models for Improved Stance Detection. In: Proceedings of the 3rd NLP4IF Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda. Barcelona, Spain: International Committee on Computational Linguistics (ICCL); 2020. p. 22–32 https://aclanthology.org/2020.nlp4if-1.3.
54. Mohammad S. Word Affect Intensities. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan: European Language Resources Association (ELRA); 2018. https://aclanthology.org/L18-1027.
55. Strapparava C, Valitutti A. WordNet-Affect: An Affective Extension of WordNet. In: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004). Lisbon: European Language Resources Association (ELRA); 2004. p. 1083–6.
56. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825–30.
57. Babanejad N, Davoudi H, An A, Papagelis M. Affective and Contextual Embedding for Sarcasm Detection. In: Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain: International Committee on Computational Linguistics; 2020. p. 225–43. https://doi.org/10.18653/v1/2020.coling-main.20.
58. Miller GA. WordNet: a lexical database for English. Commun ACM. 1995;38(11):39–41. https://doi.org/10.1145/219717.219748.
59. Brown-Johnson CG, Cataldo PhD JK, Orozco N, Lisha NE, Hickman N, Prochaska JJ. Validity and Reliability of the Internalized Stigma of Smoking Inventory: An Exploration of Shame, Isolation, and Discrimination in Smokers with Mental Health Diagnoses. Am J Addict. 2015;24(5):410–8. https://doi.org/10.1111/ajad.12215.
60. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Wallach H, Larochelle H, Beygelzimer A, Alché-Buc F d', Fox E, Garnett R, editors. Advances in Neural Information Processing Systems 32. Curran Associates, Inc.; 2019. p. 8024–35. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.
61. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. 2019. https://doi.org/10.48550/arXiv.1910.03771.
62. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika. 1947;12(2):153–7. https://doi.org/10.1007/BF02295996.
63. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J Roy Stat Soc: Ser B (Methodol). 1995;57(1):289–300.
64. Dror R, Baumer G, Shlomov S, Reichart R. The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics; 2018. p. 1383–92. https://doi.org/10.18653/v1/P18-1128.
65. Creswell JW, Clark VLP. Designing and Conducting Mixed Methods Research. SAGE Publications; 2017.
66. Doyle L, Brady AM, Byrne G. An overview of mixed methods research. J Res Nurs. 2009;14(2):175–85. https://doi.org/10.1177/1744987108093962.
67. Agresti A. Categorical data analysis, Second Edition. New York: Wiley; 2002.
68. Sharpe D. Chi-square test is statistically significant: Now what? Pract Assess Res Eval. 2015;20(1):8.
69. Moreno MA, Goniu N, Moreno PS, Diekema D. Ethics of Social Media Research: Common Concerns and Practical Considerations. Cyberpsychol Behav Soc Netw. 2013;16(9):708–13. https://doi.org/10.1089/cyber.2012.0334.
70. Galinsky AD, Wang CS, Whitson JA, Anicich EM, Hugenberg K, Bodenhausen GV. The Reappropriation of Stigmatizing Labels: The Reciprocal Relationship Between Power and Self-Labeling. Psychol Sci. 2013;24(10):2020–9. https://doi.org/10.1177/0956797613482943.
71. Jawahar G, Sagot B, Seddah D. What Does BERT Learn about the Structure of Language? In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics; 2019. p. 3651–7. https://doi.org/10.18653/v1/P19-1356.
72. Khetan V, Ramnani R, Anand M, Sengupta S, Fano AE. Causal BERT: Language Models for Causality Detection Between Events Expressed in Text. In: Arai K, editor. Intelligent Computing. Cham: Springer International Publishing; 2022. p. 965–80. (Lecture Notes in Networks and Systems). https://doi.org/10.1007/978-3-030-80119-9_64.
73. Kacewicz E, Pennebaker JW, Davis M, Jeon M, Graesser AC. Pronoun Use Reflects Standings in Social Hierarchies. J Lang Soc Psychol. 2014;33(2):125–43. https://doi.org/10.1177/0261927X13502654.
74. Cohn MA, Mehl MR, Pennebaker JW. Linguistic Markers of Psychological Change Surrounding September 11, 2001. Psychol Sci. 2004;15(10):687–93. https://doi.org/10.1111/j.0956-7976.2004.00741.x.
75. Moore MD, Ali S, Burnich-Line D, Gonzales W, Stanton MV. Stigma, Opioids, and Public Health Messaging: The Need to Disentangle Behavior From Identity. Am J Public Health. 2020;110(6):807–10. https://doi.org/10.2105/AJPH.2020.305628.
76. O'Connor C, McNamara N, O'Hara L, McNicholas M, McNicholas F. How do people with eating disorders experience the stigma associated with their condition? A mixed-methods systematic review. J Ment Health. 2021;30(4):454–69. https://doi.org/10.1080/09638237.2019.1685081.

Publisher's Note