The LGBTQ+ Minority Stress on Social Media (MiSSoM) Dataset: A Labeled Dataset for Natural Language Processing and Machine Learning

Cory J. Cascalheira[1], Santosh Chapagain[2], Ryan E. Flinn[3], Dannie Klooster[4], Danica Laprade[5], Yuxuan Zhao[1], Emily M. Lund[6], Alejandra Gonzalez[7], Kelsey Corro[1], Rikki Wheatley[8], Ana Gutierrez[1], Oziel Garcia Villanueva[1], Koustuv Saha[9], Munmun De Choudhury[10], Jillian R. Scheer[11], Shah M. Hamdi[2]

[1]New Mexico State University,
[2]Utah State University,
[3]University of North Dakota,
[4]Oklahoma State University,
[5]Northern Arizona University,
[6]University of Alabama,
[7]Xavier University,
[8]University of Oregon,
[9]University of Illinois Urbana-Champaign,
[10]Georgia Tech,
[11]Syracuse University
cjcascalheira@gmail.com, santosh.chapagain@usu.edu, ryanelliottflinn@gmail.com, dannie.klooster@okstate.edu, drl276@nau.edu, zyx8010@nmsu.edu, emlund@ua.edu, gonzaleza22@xavier.edu, kcorro@nmsu.edu, rikkiw@uoregon.edu, anagut22@nmsu.edu, ozielw@nmsu.edu, koustuv.saha@gmail.com, munmun.choudhury@cc.gatech.edu, jrscheer@syr.edu, s.hamdi@usu.edu

Abstract

Minority stress is the leading theoretical construct for understanding LGBTQ+ health disparities. As such, there is an urgent need to develop innovative policies and technologies to reduce minority stress. To spur technological innovation, we created the largest labeled datasets on minority stress using natural language from subreddits related to sexual and gender minority people. A team of mental health clinicians, LGBTQ+ health experts, and computer scientists developed two datasets: (1) the publicly available LGBTQ+ **Mi**nority **S**tress on **So**cial **M**edia (MiSSoM) dataset and (2) the advanced request-only version of the dataset, LGBTQ+ MiSSoM+. Both datasets have seven labels related to minority stress, including an overall composite label and six sublabels. LGBTQ+ MiSSoM (N = 27,709) includes both human- and machine-annotated labels and comes preprocessed with features (e.g., topic models, psycholinguistic attributes, sentiment, clinical keywords, word embeddings, n-grams, lexicons). LGBTQ+ MiSSoM+ includes all the characteristics of the open-access dataset, but also includes the original Reddit text and sentence-level labeling for a subset of posts (N = 5,772). Benchmark supervised machine learning analyses revealed that features of the LGBTQ+ MiSSoM datasets can predict overall minority stress quite well (F1 = 0.869). Benchmark performance metrics yielded in the prediction of the other labels, namely prejudiced events (F1 = 0.942), expected rejection (F1 = 0.964), internalized stigma (F1 = 0.952), identity concealment (F1 = 0.971), gender dysphoria (F1 = 0.947), and minority coping (F1 = 0.917), were excellent. Descriptive analyses, ethical considerations, limitations, and possible use cases are provided.

Introduction

Minority stress is the psychosocial strain experienced by lesbian, gay, bisexual, transgender, queer, and other sexual and gender minority (LGBTQ+) people who are exposed to anti-LGBTQ+ stigma (Brooks 1981; Meyer 2003). Across the fields of social science, medicine, and public health, minority stress is the leading theoretical construct explaining health disparities between LGBTQ+ people and their cisgender and heterosexual peers (Institute of Medicine of the National Academies 2011; Pachankis et al. 2020; de Lange et al. 2022). As such, there is an urgent need to develop innovative policies and technologies to better understand, and ultimately reduce, reduce minority stress.

Minority stress is a composite phenomenon (Cascalheira

et al. 2023a)—that is, overall minority stress is composed of *factors of minority stress* (Meyer 2003; Lindley and Galupo 2020), namely prejudiced events (i.e., discrimination), identity concealment (i.e., hiding one's LGBTQ+ identity), expected rejection (i.e., imagining one will face prejudice or discrimination), internalized stigma (i.e., believing anti-LGBTQ+ ideas as if they were true), and gender dysphoria (i.e., incongruence between one's gender identity and societal expectations for one's assigned sex at birth).

Contributions of the Present Paper

While health and social science researchers have prioritized the concept of minority stress as a means of understanding LGBTQ+ health, the field of artificial intelligence (AI) has yet to embrace its utility (Saha et al. 2019; Cascalheira et al. 2022, 2023b, a). This failure is, in part, a result of limited access to robust data sets that specifically capture LGBTQ+ minority stress in a manner that is scientifically rigorous and ethically sound. Thus, this group has documented the creation of such a dataset using a set of standardized guidelines (Gebru et al. 2021). This paper makes several substantial contributions to the field, including:

- Introduction and explanation of the LGBTQ+ Minority Stress on Social Media (MiSSoM) datasets—the largest labeled datasets on minority stress. These data sets use natural language from Reddit.com, and include open access and restricted access versions.
- Overview of exploratory analyses describing the language of minority stress on social media.
- Training of four supervised machine learning models and one neural network on the open-access LGBTQ+ MiSSoM features, providing benchmark performance metrics to other researchers.

Motivation

While the creation of these datasets provides AI scientists with robust information that is scientifically rigorous and ethically sound for the study of minority stress on social media, the ultimate goal of their creation is to offset the effects of minority stress in the lives of LGBTQ+ individuals by leading to the development of interventions and practices that promote LGBTQ+ health equity. To this end, the LGBTQ+ MiSSoM datasets facilitate expansion of minority stress theory through natural language processing investigations of community-driven (vs. theory-driven) language. Minority stress is a linguistically sophisticated health construct (Cascalheira et al. 2023a), and its nuances are not fully understood. These datasets provide an opportunity for researchers to use computational social science to compare expressions of minority stress across posts, explore the motivation of minority stress disclosures online, and understand LGBTQ+ help-seeking for minority stress. These datasets can also help reduce minority stress in LGBTQ+ individuals through the development of AI-enhanced digital health interventions (e.g., training AI models with these data and employing them in stress-detection systems). Additionally, they have utility related to:

- Sentence-level prediction of minority stress in social media (which could be useful in fine-grained linguistic analysis using sentence-embedding algorithms);
- Prediction of LGBTQ+ health outcomes (e.g., AI studies of depression among LGBTQ+ people on social media) using the LGBTQ+ MiSSoM datasets as features;
- Data collection related to public health interventions, such as surveilling surges in minority stress following the passage of anti-LGBTQ+ laws (Cascalheira et al. 2023a);
- Instruction of scholars and students interested in LGBTQ+ health on methodology to train AI models.

Dataset Description

This paper introduces two tabular datasets where each row represents an individual social media post. First, we provide the publicly available LGBTQ+ MiSSoM dataset (N = 27,709), an NLP dataset created from Reddit.com posts and comments. The LGBTQ+ MiSSoM dataset includes both human- and machine-annotated labels and comes preprocessed with features (e.g., topic models, psycholinguistic attributes, sentiment, word embeddings, n-grams, lexicons). The labels cover major factors of minority stress theory, including prejudiced events, internalized stigma, identity concealment, expected rejection, and minority coping. Gender dysphoria is also included as a label given its recent theoretical framing as a proximal stressor. Second, we provide the restricted, request-only version of the dataset—LGBTQ+ MiSSoM+, which includes everything in the public version plus the original Reddit.com text, names of the subreddits, and sentence-level labeling for 5,789 posts and comments. Separate datasets with different access levels were developed to protect the LGBTQ+ community (see Ethical Considerations). Both datasets are available from the Harvard Dataverse (for hyperlinks, see FAIR Data Principles).

Data Collection

As shown in Fig. 1, data were scraped from LGBTQ+-related subreddits on Reddit.com using PushShift (Baumgartner et al. 2020). Subreddits were selected to capture linguistic content from LGBTQ+ people with different sexual orientations and gender identities. Data from r/lgbt were not downloaded because a previous paper had already created an LGBTQ+-themed dataset with r/lgbt (Saha et al. 2019).

Using PushShift, we scraped the first 10,000 posts from each subreddit beginning on 18 September 2021. Some posts were scraped from as far back as 18 August 2012. Empty, deleted, and removed posts were eliminated, yielding a dataset of 27,796 posts. Figure 1 shows the proportion

of posts yielded by each subreddit.

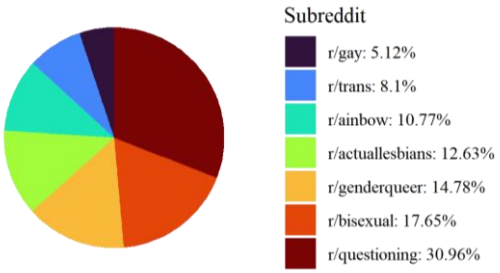

Figure 1: Subreddits Included in the MiSSoM Datasets.

Human Annotation

We used an iterative process of consensus-based, conventional content analysis (Hsieh and Shannon 2005). Our coding team included nine LGBTQ+ health experts, clinicians, students, and community members diverse in age, race, ethnicity, sexual orientation, gender identity, and disability status. Informed by best practices for increasing the trustworthiness of the ground truth labels (Morrow 2005; Creswell and Poth 2018), our team (a) generated a robust codebook informed by minority stress theory, borrowed from a previous study (Saha et al. 2019), and modified by empirical examples; (b) underwent an extensive training period to achieve consensus on the codebook and boundaries of the labels (i.e., Cohen's $\kappa \geq 0.80$ or simple percent agreement ≥ 0.90; O'Connor and Joffe 2020); and (c) met weekly between September 2021 and February 2023 to resolve disagreements between coders, iteratively modify the codebook, track major coding decisions, continue discussions about limitations and boundaries of the annotation process, and audit one another's annotation decisions. Additionally, the first author randomly audited 25 annotations each week. We used the labeling software Tagtog to complete annotations. The team coded 5,789 Reddt.com posts and comments; after preprocessing (see Preprocessing and Feature Generation), a total of 5,772 posts and comments with human-annotated labels remained. The codebook used in the present project is available on the Open Science Framework: https://osf.io/qgah8/

Seven Labels of Minority Stress

A total of 5,789 Reddit.com posts and comments were annotated by humans. Annotation yielded seven labels related to minority stress, plus a negative label (i.e., the absence of minority stress). Table 1 shows text examples of each label; the examples are slightly paraphrased and edited to protect the privacy of the Reddit.com user. The overall, composite label of minority stress was present if at least one factor of minority stress other than gender dysphoria (i.e., prejudiced events, internalized stigma, expected rejection, identity concealment) was coded by a member of the annotation team. Gender dysphoria was not part of the composite label due to its newer, and still tentative (Lindley and Galupo 2020), addition to the minority stress model but was analyzed in a separate article (not sure if we need the citation to be anonymous here?). Minority coping was coded if a factor of minority stress was present *and* there was evidence of coping as defined by Meyer (2003). We decided to code minority coping in this manner because, without such a rule, one might argue many (if not most) Reddit.com posts and comments on LGBTQ+-themed subreddits function as a form of community-based coping via social support seeking.

Construct	Label	Text
Minority Stress (Composite)	1	I'm doomed to be anything but cis […] nobody gets me […] they all call it a phase […]
Prejudiced Event	1	my gf's family is homophobic […] I feel like shit over it […] her family doesn't understand […] they see me as some dyke who is ruining my gf's life
Expected Rejection	1	i came out as bisexual to my fam but my best bud doesnt know […] i'm worried about what hes gonna say
Internalized Stigma	1	I'm not the best at being gay, and still feel a lot of shame
Identity Concealment	1	I'll be honest [I'm gay] but I haven't really told anyone
Minority Coping	1	[my grandmother] might not understand if I come out […] has anybody been in a similar situation? What did you […]
Gender Dysphoria	1	i have some intense dysphoria swings […]
Negative Label	0	[my friend] recently came out […] they asked that I talk to them using different pronouns

Table 1. Text Examples of Each Label.

Major Annotation Decisions

Throughout the iterative process of consensus-based content analysis, our team discussed points of disagreement on the boundaries of coding minority stress on social media. We made several important decisions, which influenced annotation and codebook development.

First, we decided to code posts at the semantic level since follow-up with the Reddit users was neither feasible nor ethical. If the intended meaning of a post could be interpreted in multiple ways, we coded based on the most explicit and literal interpretation.

Second, we differentiated between minority stressors and general stressors. If distress was present but was not explicitly attributed by the LGBTQ+ person to a minority stress experience (e.g., internalized stigma), and/or could be reasonably expected to result from another experience also mentioned in the post (e.g., sexual assault), we erred on the side of not coding material as minority stress.

Third, we coded references to "homophobic" or "transphobic" people as expected rejection. During peer debriefing, we noticed instances in which LGBTQ+ people mentioned the presence of "homophobic/transphobic" people in their lives yet did not elaborate on the specific actions of these individuals (i.e., whether they committed a prejudiced action) or the impact of the person's actions (i.e., whether the homophobia/transphobia became internalized). We determined that amorphous references to homophobia, transphobia, and related hate (e.g., biphobia) functioned to communicate *potential* prejudice, and therefore were semantic indicators of expected rejection.

Fourth, we focused on *present* distress and impairment. A focus on present tense builds upon past research on minority stress on social media (Saha et al. 2019). A focus on distress and impairment is consistent with clinical guidelines for determining significant distress (American Psychiatric Association 2013). Several LGBTQ+ people described experiences of minority stress that bothered them previously but no longer caused distress (e.g., telling a story about being in the closet for years and ending with a celebratory comment about coming out). When LGBTQ+ people wrote about experiences of minority stress from the past without indicating whether these experiences continued to cause distress or impairment presently, we did not annotate the post as an example of minority stress.

Finally, we realized after scraping Reddit and beginning to annotate the posts that LGBTQ+ people would frequently allude to material they had written in the post title, and in the absence of the post's title (which was decoupled from the body of the post in the downloading process) the intended meaning could not be determined. Even when we suspected minority stress was present, if minority stress was not clearly present without the context provided by the post title, we did not annotate the presence of minority stress.

Limitations of Annotation

Several posts contained indicators of stress and coping which we judged as either not consistent with core theoretical constructs of LGBTQ+ minority stress or too ambiguous to annotate with certainty. These posts, or portions of posts, were annotated as "limitations of the codebook" to inform further research in subsequent ML- and NLP-based studies of LGBTQ+ minority stress. Although we generated two dozen specific limitations during codebook developed, we thematically analyzed these limitations to identify six primary categories of excluded material:

1. stress resulting from stigmatization of relationship configuration diversity (e.g., polyamory, consensual non-monogamy, asexual and/or aromantic relationships);
2. intersectional stigma in which LGBTQ+ minority stress was not the primary cause of stress, LGBTQ+ minority stress intersected substantially with other forms of oppression (e.g., racism, sizeism, classism, ableism), or experiences of LGBTQ+ minority stress enacted by other members of the LGBTQ+ community (e.g., biphobia within the LGBTQ+ community);
3. stress communicated in a language other than English, clearly generated by a bot, or written with extensive self-censoring (e.g., substituting letters with asterisks), which rendered the meaning ambiguous;
4. stressful experiences communicated through poems, song lyrics, or other forms of art as the attribution of such experiences to the user was ambiguous;
5. environmental or structural discrimination without a clear impact on the LGBTQ+ person (e.g., pondering a new anti-transgender law); and
6. microaggressions (Nadal 2019) wherein it was unclear if the LGBTQ+ person perceived the experience as discriminatory or prejudiced.

Machine Annotation

A total of 22,007 Reddit.com posts and comments were annotated by a hybrid neural network, Bidirectional Encoder Representations from Transformers convolutional neural network (BERT-CNN); after preprocessing the data, a total of 21,937 posts and comments with machine-annotated labels remained. The training and selection of BERT-CNN is described fully elsewhere (Cascalheira et al. 2023a). Briefly, each Reddit.com post was tokenized, embedded with BERT, fed into a convolutional layer with max pooling, corrected with a dropout layer, and converted to a fully connected neural network before labels were predicted with the sigmoid function. BERT-CNN was used to machine-annotate the each post due to its excellent performance in classifying both composite minority stress (F1 = 0.84) and individual factors of minority stress—namely, prejudiced events (F1 = 0.87), expected rejection (F1 = 0.92), internalized stigma (F1 = 0.91), identity concealment (F1 = 0.92), minority coping (F1 = 0.84), gender dysphoria (F1 = 0.94).

Preprocessing and Feature Generation

Prior to engineering features, raw text was preprocessed by removing URLs, excessive whitespace, and recoding special characters (e.g., "&" → "and"). For all features except for the psycholinguistic attributes and word embeddings, text was lowercased, stop words were removed, contractions were expanded, and lemmatization was performed. Raw text was used for the psycholinguistic attributes and word embeddings because each algorithm makes use of all linguistic content (Mikolov et al. 2013; Pennebaker et al. 2015).

After all features and labels were created, the datasets were audited for missing values. Only 87 cases with missing values were found. Thus, all cases with missing values were eliminated with case-wise deletion.

A total of 726 features were engineered across 10 categories. Feature categories are organized in order of computational complexity. Unless otherwise stated, string detection and regular expressions were used to create the features.

Clinical Keywords

Medicalized language used to describe mental and behavioral health disorders may be useful signals given the LGBTQ+ community's health inequities (Goldbach et al. 2014; Pachankis et al. 2020; Cascalheira et al. 2023b). Thus, text from the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-5*) chapters on anxiety, depression, stress disorders, substance use, and gender dysphoria were mined for the most frequent 10 clinical keywords. Fifteen transdiagnostic clinical keywords that appeared across *DSM-5* chapters (e.g., "diagnosis") were grouped into a separate feature. If at least one clinical keyword was present, then the feature was assigned 1, otherwise 0. Six individual features were created.

n-Grams

We generated unigrams (n = 100), bigrams (n = 100), and trigrams (n = 100) after data preprocessing by calculating the term-frequency inverse document frequency (TF-IDF) scores to extract the top 300 n-Grams. The first author conducted a close inspection of each list using domain knowledge to remove common but noisy words (e.g., days of the week), pandemic-related words (e.g., COVID-19), and nonsense words (e.g., amp) to increase the probability of the features being strong signals of minority stress. N-Grams were coded as present (1) or absent (0).

Sentiment Lexicon

Sentiment is widely used in natural language processing to predict constructs, so a continuous feature assessing positive and negative sentiment was created using AFINN (Finn 2011) and slangSD (Wu et al. 2018) lexicons. Each lexicon assigns a positive or negative value to each word (e.g., "adulting" = −1, "yummy" = +3).; words without a lexicon match were assigned zero. The summation of sentiment of each post was calculated such that higher positive values correspond to greater positive sentiment, lower negative values indicate greater negative sentiment, and values close to zero represent an overall neural or balanced sentiment.

Hate Speech Lexicon

Minority stress encompasses prejudiced events that are directed towards SGM people (Meyer 2003), such as verbal violence (e.g., using offensive slurs; Saha et al., 2019). Thus, the Hatebase (2022) lexicon was used with string detection and regular expressions to identify posts exhibiting English hate speech terms related to sexual minority (n = 68) and gender minority (n = 77) groups. Hate speech terms were coded as present (1) or absent (0).

Minority Stress Theoretical Lexicon

Because LGBTQ+ people often use clinical words to describe their experiences with gender dysphoria on social media (Cascalheira et al. 2023b), and considering the pervasiveness of minority stress (Meyer 2003), it is possible that this group uses other jargon-heavy words to describe their experiences with minority stress. Thus, several seminal papers on minority stress (Meyer 1995, 2003; Balsam et al. 2011; Hendricks and Testa 2012) were mined for the most frequent keywords. Common words were removed (e.g., "gay", "measure"). The first author closely inspected each word using domain knowledge. A single feature was created indicating whether terms were present (1) or absent (0).

Pain Lexicon

Because stress has a biological component (Epel et al. 2018), LGBTQ+ adults may talk about their psychophysiological reactions to minority stress (e.g., "discomfort"). Hence, a pain lexicon was used (Chaturvedi et al. 2021). The pain lexicon was developed using symptoms from patient-authored text, words relating to biomedical terms for pain, synonyms for pain from biomedical ontologies (e.g., The Unified Medical Language System), and word embeddings to find words that were similar to pain in latent lexicosemantic space. The pain lexicon contains 382 terms related to psychophysiological complaints (Chaturvedi et al. 2021). These terms were detected with regular expressions to determine if the psychophysiological complaint was present (1) or absent (0).

Psycholinguistic Attributes

Psycholinguistic attributes have been shown to be strong signals of mental health constructs on social media (Coppersmith et al. 2014), including minority stress (Saha et al. 2019). Psycholinguistic attributes were generated with the Linguistic Inquiry and Word Count (LIWC) algorithm (Pennebaker et al. 2015). LIWC computes 93 features spanning word count, summary variables (e.g., words with at least six letters), linguistic dimensions (e.g., first person pronoun use), grammar use (e.g., if verbs are used), and psychological processes (e.g., cognition, social affiliation).

Topic Models

Topic models are another form of open vocabulary that can describe constructs on social media (Schwartz et al. 2013). Latent Dirichlet analysis was performed with a limit of 50 topics; inspection of an elbow plot visualizing the UMass coherence score and the number of topics was used to select 10 topics as optimal (Blei et al. 2003). Each topic was added as a feature with values indicating its presence (1) or absence (0) in each Reddit.com post/comment.

Word Embeddings

Past research indicates that word embeddings are predictive of minority stress disclosures on social media (Saha et al 2019). Thus, we used Word2Vec (Mikolov et al. 2013) to create 300 features, taking the summation of each individual

dimension of the Word2Vec model to create the features.

FAIR Data Principles

LGBTQ+ MiSSoM and MiSSoM+ are findable, accessible, interoperable, and reusable (FAIR). Both datasets are hosted on the Harvard Dataverse. All scripts used in the present paper are available on GitHub.

- https://doi.org/10.7910/DVN/GPRSXH
- https://github.com/CJCascalheira/lgbtq_stress_dataset

Data are findable on the Harvard Dataverse with important keywords. The data have rich metadata and specify the data record identifier.

Data are accessible using the Harvard Dataverse. The metadata are always visible. Clear instructions on accessing LGBTQ+ MiSSoM and MiSSoM+ are provided. Specifically, anyone may download LGBTQ+ MiSSoM. To access LGBTQ+ MiSSoM+, qualified researchers must use the request access feature on the Harvard Dataverse. Researchers must disclose conflicts of interests, state their purpose for using the data, agree to keep the dataset off publicly accessible web spaces and not share the data, agree to not sell the data, describe a data protection plan, list all entities who will use the data, and provide ethical approval (e.g., IRB) for research purposes. LGBTQ+ MiSSoM+ access is managed by the first author. Separating the LGBTQ+ MiSSoM and MiSSoM+ by access levels preserves community standards of safety and justice among LGBTQ+ people.

Finally, data are interoperable and reusable. Existing studies using the LGBTQ+ MiSSoM+ dataset are cataloged on the Harvard Dataverse. All language is written in an accessible manner. Both the LGBTQ+ MiSSoM and MiSSoM+ are licensed under Creative Commons Attribution 4.0 International (CC BY 4.0).

Exploratory Data Analysis

Exploratory data analyses were conducted to contextualize the LGBTQ+ MiSSoM datasets. Figures 3 and 4 show word clouds created using TF-IDF scores where the size of the word represents its prevalence in the data. The clearest difference between the positive and negative labels of composite minority stress are references to time and feeling/thinking, respectively. Identity labels were used more in positive examples of minority stress, whereas relationship language was used more in negative examples of minority stress.

Figure 3: Word Cloud of Positive Class

Figure 4: Word Cloud of Negative Class

Differences in Minority Stress Labels

We conducted *t*-tests with the Bonferroni correction to examine differences in psycholinguistic attributes between the composite minority stress positive and negative label. As shown in Table 2, in which significantly different psycholinguistic attributes are organized by Cohen's *d* (i.e., a measure of effect size), several significant differences emerged. Note that some variables tested with low Cohen's *d* are not depicted in Table 2 due to space limitations, but can be retrieved from the GitHub code. In terms of meaningfully significant differences, positive examples of minority stress tended to have higher word counts, discuss family experiences more, reference the self more, and exhibit genuineness more than Reddit.com posts and comments without evidence of minority stress. Negative examples, in contrast, tended to use an uplifting tone, focus on intellectual discussions, discuss the Reddit users in self-enhancing ways, and ask questions. Reddit users classified as reporting minority stress referred to people in their lives more, told stories anchored in the past, referenced their home environment, and discussed negative emotions like anxiety. Reddit users who did not exhibit minority stress in their posts and comments were more informal, used internet idioms, and evinced greater positive affect.

Feature	*t*	*D*	Greater Mean
WC	-14.344	0.526	Positive Class
Family	-12.868	0.463	Positive Class
I	-15.302	0.404	Positive Class
Authentic	-14.189	0.38	Positive Class
Tone	12.441	0.375	Negative Class
Function	-16.782	0.358	Positive Class
Analytic	12.784	0.329	Negative Class
Pronoun	-12.734	0.323	Positive Class
Clout	11.17	0.318	Negative Class
Ppron	-12.33	0.317	Positive Class
Home	-9.143	0.309	Positive Class
Anx	-9.145	0.265	Positive Class
Dic	-13.274	0.255	Positive Class
QMark	12.322	0.239	Negative Class
Negemo	-9.662	0.238	Positive Class
Verb	-9.251	0.233	Positive Class
posemo	9.934	0.224	Negative Class
negate	-8.395	0.218	Positive Class
WPS	-6.57	0.211	Positive Class
focuspast	-6.78	0.2	Positive Class
prep	-7.557	0.191	Positive Class
netspeak	8.799	0.169	Negative Class
you	6.685	0.163	Negative Class
informal	8.193	0.159	Negative Class
risk	-5.402	0.158	Positive Class
AllPunc	7.69	0.152	Negative Class
OtherP	7.888	0.145	Negative Class
motion	-5.292	0.143	Positive Class
anger	-5.364	0.141	Positive Class
conj	-5.692	0.141	Positive Class
number	7.171	0.141	Negative Class
see	6.234	0.139	Negative Class
Apostro	-4.76	0.129	Positive Class
tentat	5.103	0.122	Negative Class
ingest	5.735	0.119	Negative Class
auxverb	-4.644	0.116	Positive Class
bio	4.793	0.114	Negative Class
they	-3.922	0.114	Positive Class
adverb	-4.445	0.108	Positive Class
leisure	4.967	0.106	Negative Class
sad	-4.006	0.105	Positive Class
article	4.121	0.103	Negative Class
focusfuture	-3.874	0.102	Positive Class
cause	-3.932	0.101	Positive Class
affiliation	4.161	0.101	Negative Class
relig	-3.544	0.099	Positive Class
we	3.617	0.094	Negative Class
body	3.925	0.092	Negative Class

Table 2. Significant Differences in Psycholinguistic Attributes between Posts

Benchmark Performance of Supervised Machine Learning Models

Four widely used supervised machine learning models, and one neural network used in past research on LGBTQ+ minority stress (Saha et al. 2019), were trained using an 80% training and 20% testing stratified split (Raschka and Mirjalili 2019). The following models were trained using scikit-learn in Python:

- **Support vector machine** finds an optimal decision boundary by maximizing margin between support vectors; trained with regularization of 1 and RBF kernel.
- **Logistic regression** models the relationship between input variables and class probability; trained with L2.
- **Random forest** is an ensemble learning method combining decision trees; trained with baseline of 100 decision trees.
- **Adaptive boosting (AdaBoost)** is an ensemble learning algorithm that iteratively trains weak classifiers on weighted versions of data; trained with *DecisionTreeClassifier*, max depth of 1, and 50 estimators.
- **Multilayer perception** is a feed-forward neural network trained with default hyperparameters, including a hidden layer of 100, Adam optimizer, and ReLU activation.

The performance metrics of each supervised machine learning model are shown in Table 3. Boldface indicates the best value for accuracy, precision, recall, and weighted F1 for each label across algorithms. In general, support vector machine, AdaBoost, and random forest yielded the best performance metrics.

Discussion

AI scientists and health professionals now have two large natural language datasets to study minority stress on social media. The LGBTQ+ MiSSoM datasets include high-quality labels produced by best practices in qualitative coding and state-of-the-art deep learning. As shown in the descriptive analyses, positive and negative examples of minority stress are syntactically and semantically different. Future researchers can probe these initial descriptive analyses for greater nuance, perhaps examining linguistic differences among factors of minority stress (e.g., how does identity concealment differ from internalized stigma semantically?) or subsetting the datasets to examine minority stress disclosure among specific and understudied LGBTQ+ subgroups (e.g., bisexual women, aromantic nonbinary people).

Using traditional machine learning algorithms and a neural network, we demonstrated that the features of the LGBTQ+ MiSSoM datasets yield excellent performance metrics in predicting composite minority stress as well as

individual factors of minority stress. In fact, our expertly-derived features and rigorously generated ground truth labels improved the prediction of composite minority stress (Saha et al. 2019) and factors of minority stress (Cascalheira et al. 2023b), and even outperformed neural network models (Cascalheira et al. 2022, 2023a).

Ethical Considerations

Ethical concerns surrounding the collection and distribution of information regarding the lives of LGBTQ+ individuals are of utmost importance to this team. Indeed, most members of the research team proudly identify as LGBTQ+.

Ultimately, these datasets were created with the intent of providing a model for the collection and analysis of information, regarding the experiences of LGBTQ+ individuals, that expands the scientific understanding of minority stress and promotes the development of practices that improve overall well-being and quality of life for LGBTQ+ people.

To promote a culture of ethical consideration and preservation of individual dignity, the research team took the following steps during the research process. Creation of the LGBTQ+ MiSSoM datasets were guided by the ethical codes of the American Psychological Association (2017) and the Association for the Advancement of Artificial Intelligence (2019). As recommended in digital work with LGBTQ+ people (Cascalheira et al. 2023c), the research team consisted of and involved members of the LGBTQ+ community throughout the research process to informally assess acceptability. Reddit was chosen as the social media platform for this development of the datasets to capitalize on the increased anonymity of its LGBTQ+ users (Leavitt 2015). We developed two separate datasets with different access levels in an effort to reduce the likelihood of it being accessed by malicious actors.

As with all electronic data of a sensitive nature, the possibility for misuse or misrepresentation of these datasets (with or without the intent to cause harm) remains. After careful consideration, the team has determined the likelihood of these potential harms to be no greater than the inherent minimal risk for Reddit users posting on the site.

Limitations

There are several limitations of the LGBTQ+ MiSSoM datasets. First, there was an absence of hyperparameter tuning (e.g., grid search), which could have improved benchmark performance of the binary classification task. Second, all posts and comments included in these datasets are assumed to be generated by humans. We did not verify the humanity of Reddit users despite the presence of bots in previous LGBTQ+ research recruiting participants from social media platforms (Simone et al. 2023). Third, and relatedly, we assumed that Reddit users in these datasets were LGBTQ+ unless they named themselves as an ally. It was not feasible to verify each Reddit user's LGBTQ+ identity. Additionally, as noted previously, we did not have access to post titles during the coding process. Finally, although Reddit.com was an excellent starting point for dataset creation given its culture of anonymity (Leavitt 2015), these data are a product of the linguistic and cultural norms of content creation on Reddit.com, and they cannot be assumed to represent the entirety of thoughts and opinions of the LGBTQ+ population. LGBTQ+ people use different social media websites to fulfill unique, platform specific needs (Craig et al. 2021). Thus, training AI models with the MiSSoM datasets is a good start, especially considering the lack of similar datasets available, but results may not transfer easily to other forms of social media (e.g., X, Facebook) or other kinds of natural language (e.g., text messages).

Conclusion

In this paper, a team of mental health clinicians, LGBTQ+ health experts, and computer scientists developed two datasets: (1) the publicly available LGBTQ+ MiSSoM dataset and (2) the restricted, request-only version of the dataset, LGBTQ+ MiSSoM+. Both datasets have seven labels related to minority stress, including an overall composite label and six sublabels. The LGBTQ+ MiSSoM datasets are offered to the AI community in hopes of improving AI research with LGBTQ+ people, expanding research on minority stress detection, and developing interventions that target minority stress to support the positive health and well-being of LGBTQ+ individuals.

Acknowledgments

We thank Jasmine Valdez, Andre Guaderrama, Zoe Lujan, Tracie Hitter, Andres E. Perez-Rojas, and Michael T. Kalkbrenner. The study was funded by the American Psychological Association Early Graduate Research Award, Michael Sullivan Diversity Scholarship, and the Adams-Cahill Graduate Research Award. Some authors completed this manuscript while supported by the National Institutes of Health (R25GM061222; R25DA035692; R25DA037190; K01AA028239-01A1) and the National Science Foundation (Award 2153379). The views presented in this article are the sole responsibility of the authors.

Algorithm	Label	Accuracy	Precision	Recall	F1 (Weighted)
Support Vector Machine	Minority Coping	**0.936**	0.908	**0.936**	0.906
	Prejudiced Event	0.948	0.936	0.948	0.926
	Expected Rejection	**0.972**	**0.962**	**0.972**	0.959
	Identity Concealment	0.980	0.961	0.980	**0.971**
	Internalized Stigma	0.963	0.941	0.963	0.946
	Gender Dysphoria	0.946	0.939	0.946	0.940
	Minority Stress	**0.874**	0.859	**0.874**	0.851
Logistic Regression	Minority Coping	0.935	0.891	0.935	0.905
	Prejudiced Event	0.946	0.925	0.946	0.929
	Expected Rejection	0.972	0.954	**0.972**	0.959
	Identity Concealment	0.979	0.961	0.979	0.970
	Internalized Stigma	0.963	0.928	0.963	0.946
	Gender Dysphoria	0.936	0.921	0.936	0.923
	Minority Stress	0.863	0.842	0.863	0.839
Random Forest	Minority Coping	**0.936**	**0.940**	**0.936**	0.905
	Prejudiced Event	0.947	**0.950**	0.947	0.921
	Expected Rejection	0.972	0.945	**0.972**	0.958
	Identity Concealment	**0.981**	0.961	**0.981**	**0.971**
	Internalized Stigma	**0.964**	0.928	**0.964**	0.946
	Gender Dysphoria	0.937	0.936	0.937	0.913
	Minority Stress	0.864	0.862	0.864	0.821
AdaBoost	Minority Coping	0.928	0.905	0.928	0.913
	Prejudiced Event	**0.949**	0.939	**0.949**	**0.942**
	Expected Rejection	0.971	0.961	0.971	**0.964**
	Identity Concealment	0.977	**0.967**	0.977	**0.971**
	Internalized Stigma	0.960	0.942	0.960	0.948
	Gender Dysphoria	**0.950**	**0.945**	**0.950**	**0.947**
	Minority Stress	0.873	0.862	0.873	0.866
Multilayer Perceptron	Minority Coping	0.925	0.912	0.925	**0.917**
	Prejudiced Event	0.941	0.936	0.941	0.938
	Expected Rejection	0.965	0.956	0.965	0.960
	Identity Concealment	0.974	0.966	0.974	0.970
	Internalized Stigma	0.959	**0.948**	0.959	**0.952**
	Gender Dysphoria	0.939	0.940	0.939	0.940
	Minority Stress	0.873	**0.866**	0.873	**0.869**

Table 3. Performance Metrics of Supervised Machine Learning

References

American Psychiatric Association. 2013. Diagnostic and statistical manual of mental disorders, 5th edn. American Psychiatric Association, Washington, DC

American Psychological Association. 2017. Ethical principles of psychologists and code of conduct. https://www.apa.org/ethics/code/index. Accessed 30 Sep 2019

Association for the Advancement of Artificial Intelligence. 2019. AAAI Code of Ethics and Professional Conduct. In: AAAI. https://www.aaai.org/Conferences/code-of-ethics-and-conduct.php

Balsam KF, Molina Y, Beadnell B, et al. 2011. Measuring multiple minority stress: the LGBT People of Color Microaggressions Scale. Cultur Divers Ethnic Minor Psychol 17:163–174. https://doi.org/10.1037/a0023244

Baumgartner J, Zannettou S, Keegan B, et al. 2020. The Pushshift Reddit dataset

Blei DM, Ng AY, Jordan MI. 2003. Latent Dirichlet allocation. J Mach Learn Res 3:993–1022. https://doi.org/10.5555/944919.944937

Brooks VR. 1981. Minority stress and lesbian women. Lexington Books, Lexington, MA

Cascalheira CJ, Chapagain S, Flinn RE, et al. 2023a. Predicting linguistically sophisticated social determinants of health disparities with neural networks: The case of LGBTQ+ minority stress. IEEE, Sorrento, Italy

Cascalheira CJ, Flinn RE, Zhao Y, et al. 2023b. Models of gender dysphoria using social media data for use in technology-delivered interventions: Machine learning and natural language processing validation study. JMIR Form Res 7:e47256. https://doi.org/10.2196/47256

Cascalheira CJ, Hamdi SM, Scheer JR, et al. 2022. Classifying minority stress disclosure on social media with bidirectional long short-term memory. Association for the Advancement of Artificial Intelligence, Atlanta, GA

Cascalheira CJ, Pugh T, Hong C, et al. 2023c. Developing digital health interventions for infectious diseases: Ethical considerations for sexual and gender minority adolescents and young adults. Frontiers in Reproductive Health 5:. https://doi.org/10.3389/frph.2023.1303218

Chaturvedi J, Mascio A, Velupillai SU, Roberts A. 2021. Development of a lexicon for pain. Frontiers in Digital Health 3:1–11

Coppersmith G, Harman C, Dredze M. 2014. Measuring post traumatic stress disorder in Twitter. In: Eighth International AAAI Conference on Weblogs and Social Media. pp 579–582

Craig SL, Eaton AD, McInroy LB, et al. 2021. Can social media participation enhance LGBTQ+ youth well-being? Development of the Social Media Benefits Scale. Social Media + Society 7:2056305121988931. https://doi.org/10.1177/2056305121988931

Creswell JW, Poth CN. 2018. Qualitative inquiry and research design: Choosing among five traditions, 4th edn. Sage Publications, Thousand Oaks, CA

de Lange J, Baams L, van Bergen DD, et al. 2022. Minority stress and suicidal ideation and suicide attempts among lgbt adolescents and young adults: A meta-analysis. LGBT Health 9:222–237. https://doi.org/10.1089/lgbt.2021.0106

Epel ES, Crosswell AD, Mayer SE, et al. 2018. More than a feeling: A unified view of stress measurement for population science. Frontiers in Neuroendocrinology 49:146–169. https://doi.org/10.1016/j.yfrne.2018.03.001

Finn ÅN. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In: Proceedings of the ESWC2011 Workshop on “Making Sense of Microposts”: Big things come in small packages 718 in CEUR Workshop Proceedings 93-98. Sun SITE Central Europe, Crete

Gebru T, Morgenstern J, Vecchione B, et al. 2021. Datasheets for datasets. arXiv 1–18

Goldbach JT, Tanner-Smith EE, Bagwell M, Dunlap S. 2014. Minority stress and substance use in sexual minority adolescents: A meta-analysis. Prevention Science 15:350–363. https://doi.org/10.1007/s11121-013-0393-7

Hatebase. 2022. Hatebase is a collaborative, regionalized repository of multilingual hate speech. In: Hatebase. https://hatebase.org/

Hendricks ML, Testa RJ. 2012. A conceptual framework for clinical work with transgender and gender nonconforming clients: An adaptation of the Minority Stress Model. Professional Psychology: Research and Practice 43:460–467. https://doi.org/10.1037/a0029597

Hsieh H-F, Shannon SE. 2005. Three approaches to qualitative content analysis. Qual Health Res 15:1277–1288. https://doi.org/10.1177/1049732305276687

Institute of Medicine of the National Academies. 2011. The health of lesbian, gay, bisexual, and transgender people: Building a foundation for better understanding. The National Academics Press, Washington, DC

Leavitt A. 2015. “This is a throwaway account”: Temporary technical identities and perceptions of anonymity in a massive online community. In: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing. Association for Computing Machinery, New York, NY, USA, pp 317–327

Lindley L, Galupo MP. 2020. Gender dysphoria and minority stress: Support for inclusion of gender dysphoria as a proximal stressor. Psychology of Sexual Orientation and Gender Diversity 7:265–275. https://doi.org/10.1037/sgd0000439

Meyer IH. 2003. Prejudice, social stress, and mental health in lesbian, gay, and bisexual populations: Conceptual issues and research evidence. Psychological Bulletin 129:674–697. https://doi.org/10.1037/0033-2909.129.5.674

Meyer IH. 1995. Minority stress and mental health in gay men. Journal of Health and Social Behavior 36:38–56. https://doi.org/10.2307/2137286

Mikolov T, Sutskever I, Chen K, et al. 2013. Distributed representations of words and phrases and their compositionality. In: Burges CJC, Bottou L, Welling M, et al. (eds) Advances in Neural Information Processing Systems. Curran Associates, Inc.

Morrow SL. 2005. Quality and trustworthiness in qualitative research in counseling psychology. Journal of Counseling Psychology 52:250–260. https://doi.org/10.1037/0022-0167.52.2.250

Nadal KL. 2019. A decade of microaggression research and LGBTQ communities: An introduction to the special issue. Journal of Homosexuality 66:1309–1316

O'Connor C, Joffe H. 2020. Intercoder reliability in qualitative research: Debates and practical guidelines. International Journal of Qualitative Methods 19:1–13. https://doi.org/10.1177/1609406919899220

Pachankis JE, Mahon CP, Jackson SD, et al. 2020. Sexual orientation concealment and mental health: A conceptual and meta-analytic review. Psychological Bulletin 146:831–871. https://doi.org/10.1037/bul0000271

Pennebaker JW, Boyd RL, Jordan K, Blackburn K. 2015. The development and psychometric properties of LIWC2015. In: The University of Texas at Austin. https://repositories.lib.utexas.edu/bitstream/handle/2152/31333/LIWC2015_LanguageManual.pdf

Raschka S, Mirjalili V. 2019. Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2, 3rd edn. Packt Publishing

Saha K, Kim SC, Reddy MD, et al. 2019. The language of LGBTQ+ minority stress experiences on social media. Proceedings of the ACM on Human-Computer Interaction 3:. https://doi.org/10.1145/3361108

Schwartz HA, Eichstaedt JC, Kern ML, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. PLOS ONE 8:e73791. https://doi.org/10.1371/journal.pone.0073791

Simone M, Cascalheira CJ, Pierce B. 2023. Quasi-experimental study examining the efficacy of multimodal bot screening tools and recommendations to preserve data integrity in online psychological research. American Psychologist. https://doi.org/10.1037/amp0001183

Wu L, Morstatter F, Liu H. 2018. SlangSD: Building, expanding and using a sentiment dictionary of slang words for short-text sentiment classification. Language Resources and Evaluation 52:839–852. https://doi.org/10.5555/3270332.3270380

Paper Checklist

(a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? YES

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

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

(d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? NO

(e) Did you describe the limitations of your work? YES

(f) Did you discuss any potential negative societal impacts of your work? YES

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

(h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data

and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? YES

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

2. Additionally, if your study involves hypotheses testing...
(a) Did you clearly state the assumptions underlying all theoretical results? N/A

(b) Have you provided justifications for all theoretical results? N/A

(c) Did you discuss competing hypotheses or theories that

might challenge or complement your theoretical results? N/A

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? N/A

(e) Did you address potential biases or limitations in your theoretical framework? N/A

(f) Have you related your theoretical results to the existing literature in social science? N/A

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? N/A

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoret-

ical results? N/A

(b) Did you include complete proofs of all theoretical results? N/A

4. Additionally, if you ran machine learning experiments...
(a) Did you include the code, data, and instructions

needed to reproduce the main experimental results (either in the supplemental material or as a URL)? YES

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? YES

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NO

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NO

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? YES

(f) Do you discuss what is “the cost“ of misclassification and fault (in)tolerance? YES

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, without compromising anonymity...

(a) If your work uses existing assets, did you cite the creators? N/A

(b) Did you mention the license of the assets? N/A

(c) Did you include any new assets in the supplemental material or as a URL? YES

(d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? YES

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? YES

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? YES

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? YES
6. Additionally, if you used crowdsourcing or conducted research with human subjects, without compromising anonymity...

(a) Did you include the full text of instructions given to participants and screenshots? N/A

(b) Did you describe any potential participant risks, with

mentions of Institutional Review Board (IRB) approvals? YES

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? N/A

(d) Did you discuss how data is stored, shared, and deidentified? yes