# Identifying Marijuana Use Behaviors among Homeless Youth: A Machine Learning Approach

Tianjie Deng, Andrew Urbaczewski, Young Jin Lee, Anamika Barman-Adhikari, Rinku Dewri

Submitted to: JMIR AI
on: October 08, 2023

# *Table of Contents*

# Identifying Marijuana Use Behaviors among Homeless Youth: A Machine Learning Approach

Tianjie Deng[1] PhD; Andrew Urbaczewski[2] PhD; Young Jin Lee[2] PhD; Anamika Barman-Adhikari[3] PhD; Rinku Dewri[4] PhD

[1]Department of Business Information & Analytics Daniels College of Business University of Denver Denver US
[2]Graduate School of Social Work University of Denver Denver US
[3]Department of Computer Science Ritchie School of Engineering and Computer Science University of Denver Denver US

**Corresponding Author:**
Tianjie Deng PhD

## *Abstract*

**Background:** Youth experiencing homelessness (YEH) suffer from substance use problems disproportionately compared to other youth. A study found that 69% of YEH meet the criteria for dependence on at least one substance compared to 1.8% of all US adolescents. In addition, they experience major structural and social inequalities which further undermine their ability to get the care they need.

**Objective:** The goal of this study is to develop a machine learning-based framework that utilizes homeless youth's social media content (posts and interactions) to predict their substance use behaviors (i.e., the probability of using certain substances). With this framework, social workers and care providers can identify and reach out to YEH who are at a higher risk of substance use.

**Methods:** We recruited 133 homeless youth at a non-profit organization located in a city in the western United States. After obtaining their consent, we collected types of data: (1) participants' social media conversations for the past year before they were recruited; (2) we asked the participants to complete a survey on their demographic information, health conditions, sexual behaviors, and substance usage behaviors. Building on the social sharing of emotions theory and social support theory, we identified important features that can potentially predict substance use. Then we used natural language processing techniques to extract such features from social media conversations and reactions and built a series of machine learning models to predict participants' marijuana use.

**Results:** We evaluate our models based on their predictive performance as well as their conformity to measures of fairness. Without predictive features from survey information, which may introduce gender and racial biases, our machine-learning models can reach an AUC of 0.74 and an accuracy of 0.77 using social media data only. We also evaluated the false positive rate for each gender and age segmentation.

**Conclusions:** We showed that textual interactions among YEH and their friends on social media can serve as a powerful resource to predict their substance usage. The framework we developed allows care providers to allocate resources efficiently to YEH in the greatest need while costing minimal overhead. It can be extended to analyze and predict other health-related behaviors and conditions observed in this vulnerable community.

(JMIR Preprints 08/10/2023:53488)
DOI: https://doi.org/10.2196/preprints.53488

## Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
Please make my preprint PDF available to anyone at any time (recommended).
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
✓ **No, I do not wish to publish my submitted manuscript as a preprint.**
2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?
✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

# Original Manuscript

Original Paper

# Identifying Marijuana Use Behaviors among Homeless Youth: A Machine Learning Approach

Tianjie Deng[1]*, Andrew Urbaczewski[1], Young Jin Lee[1], Anamika Barman-Adhikari[2], Rinku Dewri[3]
[1] Department of Business Information & Analytics, Daniels College of Business, University of Denver, Denver, CO 80210
[2] Graduate School of Social Work, University of Denver, Denver, CO 80210
[3] Department of Computer Science, Ritchie School of Engineering and Computer Science, University of Denver, Denver, CO 80210

*Corresponding Author:
Tianjie Deng, PhD
Department of Business Information & Analytics
Daniels College of Business
University of Denver
2101 S. University Blvd, Denver, CO 80208-8952
United States
Phone: 3038712155, Fax: 3038713695
Email: tianjie.deng@du.edu



## Abstract

**Background:** Youth experiencing homelessness (YEH) suffer from substance use problems disproportionately compared to other youth. A study found that 69% of YEH meet the criteria for dependence on at least one substance compared to 1.8% of all US adolescents. In addition, they experience major structural and social inequalities which further undermine their ability to get the care they need.
**Objective:** The goal of this study is to develop a machine learning-based framework that utilizes homeless youth's social media content (posts and interactions) to predict their substance use behaviors (i.e., the probability of using marijuana). With this framework, social workers and care providers can identify and reach out to YEH who are at a higher risk of substance use.
**Methods:** We recruited 133 homeless youth at a non-profit organization located in a city in the western United States. After obtaining their consent, we collected types of data: (1) participants' social media conversations for the past year before they were recruited; (2) we asked the participants to complete a survey on their demographic information, health conditions, sexual behaviors, and substance usage behaviors. Building on the social sharing of emotions theory and social support theory, we identified important features that can potentially predict substance use. Then we used natural language processing techniques to extract such features from social media conversations and reactions and built a series of machine learning models to predict participants' marijuana use.
**Results:** We evaluate our models based on their predictive performance as well as their conformity to measures of fairness. Without predictive features from survey information, which may introduce gender and racial biases, our machine-learning models can reach an AUC of 0.72 and an accuracy of 0.81 using social media data only when predicting marijuana use. We also evaluated the false positive rate for each gender and age segmentation.
**Conclusions:** We showed that textual interactions among YEH and their friends on social media can serve as a powerful resource to predict their substance usage. The framework we developed allows

care providers to allocate resources efficiently to YEH in the greatest need while costing minimal overhead. It can be extended to analyze and predict other health-related behaviors and conditions observed in this vulnerable community.

**Keywords:** Machine learning; Youth experiencing homelessness; Natural language processing, Infodemiology; Social good; Digital intervention

# Identifying Marijuana Use Behaviors among Homeless Youth: A Machine Learning Approach

## Introduction

> *"The drugs [f****d] meee off for awhileee but [s**t] i gotta do what i gotta do ima addict so the [f**k] what i like to get high and yaehhhh.... facebook is lameee so you probaably wont seee me on here no more... no phoneee so ill see everyone when I see themm. i am sorry i didnt say goodbye…"*
>
> - Social media posting by a homeless youth in Denver, expletives modified

## Background

The youth (persons between the ages of 18 and 24 years old) experiencing homelessness (YEH) are a vulnerable and marginalized demographic in our society. A report by the National Institute on Drug Abuse in 2019 shows that YEH engage in substantially higher levels of substance use compared to housed youth [1]. Similarly, a study found that 69% of YEH meet the criteria for dependence on at least one substance compared to 1.8% of all US adolescents [2]. The high rate of substance abuse among YEH has a variety of proven detrimental effects, including a lower level of perceived health, depression, and maladaptive coping [3], and a higher likelihood of risky sexual behaviors [4].

Despite the high rates of substance use among YEH, studies have shown that they are not receiving the intervention they need [5]. Systematic inequality exists in the distribution of such opportunities for several reasons. First, intervention facilities are not equally distributed geographically [6] - there are usually fewer non-profit facilities in economically depressed areas where YEH frequently reside. Second, the breadth and depth of intervention programs are not equally distributed among substance users. YEH have limited access to care due to their lack of insurance and financial hardship [7]. Third, intervention programs, often geared toward housed adults, are not designed equally for different substance user groups [5]. Conventional intervention services may be ineffective for YEH due to their transient lifestyle as well as structural and social barriers [7,8].

*Social media* can serve as a venue for providing efficient intervention for YEH or adolescents [9]. The example quoted earlier demonstrates that YEH may leave cues indicating their intention to initiate, quit, or relapse with substance use on their social media. The widespread and open use of social media among YEH [10] provides us with the opportunity to leverage this information and identify YEH at risk of substance use. Such identification is the first step for intervention programs tailored to YEH, which can ameliorate the inequality in the intervention process in several ways. First, social media is not constrained by location and would allow intervention programs to improve their responsiveness to patients regardless of residence. Second, integrating a social media tool into the existing intervention efforts would be less costly to implement than traditional place-based intervention programs due to lower staffing overhead and a passive collection process, allowing for more funds to be dedicated to greater *quality care* and to servicing patients experiencing financial hardship such as YEH. Third, a social media intervention tool could provide some *flexibility* that would accommodate the transient lifestyle of YEH.

## Objective

In this study, we develop a framework that can help mental health and social work professionals identify homeless youth who may be at risk of substance use through their conversations with peers

on social media. We draw from concepts in the social sharing of emotions theory [11] and social support theory [12] to develop feature sets that can be used to predict YEH's substance use behaviors. Then we built a framework that utilized a variety of natural language processing (NLP) techniques to extract such features from YEH's social media data and apply multiple machine learning (ML) models to predict their substance use behaviors [13]. The goal of these models is to provide early warnings to social workers and other health professionals so that they can prioritize and provide helpful interventions to YEH before their condition worsens or causes irreparable harm. Particularly, in this study, we will focus on predicting marijuana use. The focus on marijuana use among YEH arises from its high prevalence in this group, necessitating targeted interventions. Despite social acceptability in some areas, marijuana use among YEH poses significant health and social risks. For instance, marijuana is one of the most widely used substances among YEH. In a study conducted by Santa Maria and colleagues, 54.5% of the 66 participants reported using marijuana [14]. While decriminalization has offered some benefits, negative health outcomes have also emerged. A systematic review links marijuana use to health issues such as psychosis, mania, and suicide, as well as structural brain changes, impaired driving, and memory and learning impairment [15,16]. The wide prevalence of marijuana use and these negative consequences necessitate continued research to understand the specific factors influencing marijuana use among YEH, allowing for the development of more effective prevention and intervention strategies.

## Mine Social Media for the General Population

Social media has become a valuable source for health informatics research [17]. The field of infodemiology [18] focuses on scanning the electronic medium (e.g., the internet and social media platforms) for user-contributed health content, to improve public health [19]. Particularly, a group of scholars applies advanced analytical methods to social media data to identify users' offline health conditions and behaviors, such as depression (e.g., [20–24]), suicide intention [25], poor mental health during the pandemic [26], and substance use [27,28].

Studies closely related to our work are the detection of *substance use behavior* or mental risk behavior through mining social media data. Desrosiers et al [29] mined ethnic minority emerging male adults' text messages and Facebook messages. They found that the higher negative affect in such texts was related to a higher frequency of substance use. Owen et al. utilized language models to detect depression in users of web-based forms [30]. Tsugawa et al. used tweets to predict depression [31]. The features used include frequencies of words, the ratio of tweet topics, the ratio of positive-affect, the ratio of negative words, hourly posting frequency, tweets per day, the average number of words per tweet, overall retweet rate, overall mention rate, the ratio of tweets containing a URL, number of users following, number of users followed. Hassanpour et al [27] built a deep-learning model to predict drug usage behavior through images and text posted on Instagram. Kosinski et al [28] used the content of Facebook "likes" of a user to predict their substance use and achieved an accuracy of 0.65. Building on this work, Ding et al [32] used content someone "likes", as well as in their status updates to predict one's substance use behavior. They identified topics in such content and found that certain topics are related to substance use and alcohol use. Similarly, scholars have found that keywords and topics mentioned on social media can be related to excessive drinking, both at the country level [33] and individual level [34]. For example, Marengo et al [34] mined Facebook posts and reported a positive relationship between nightlife-related and swear-related words and problem drinking.

While the abovementioned studies incorporate offline substance behavior, they lack three aspects important to our population of interest. First, they are designed for a general population or housed youth. As we discussed earlier, substance use is much higher among YEH and follows a unique pattern that needs a targeted approach to understanding and interventions. This elevated substance use is symptomatic of the social and emotional challenges that these young people face. This usually

means that they have different needs and priorities than other young people. These needs may be reflected in their social media interactions in terms of topics or words discussed. Second, except for a few studies (e.g., [34,35]) that examined the topics discussed in people's social media posts, many of the studies aiming to predict substance use behavior utilize word embedding features that are difficult to interpret. We propose to include topic modeling in our framework so that the specific topics discussed in one's social media posts are included as an output of our framework. Third, these studies mainly focus on one's own narrative on social media and neglect the interactions they receive from their peers. Literature on substance use has suggested the importance of social support in reducing substance use [12]. Unique features that are afforded by social media platforms, such as comments, reactions, and other interactive features, may provide the social support needed by homeless youth. Studies that use social media conversations to predict other mental risk behaviors also used interaction as a predictive feature [36,37]. Therefore, features captured in such interactions should be considered.

## Using Social Media to Understand YEH's Health Behaviors

Similar to the general young adult population, YEH use social media to stay connected with their peers as well as other family members [38,39]. A series of studies assessing the social media usage pattern and sexual health behaviors among YEH help us gain some preliminary insights about how social media is associated with sexual health behaviors among this at-risk population (e.g., [40–42]). These survey-based studies report a relationship between YEH's online social networking behavior and the tendency to seek sexual-related health information and engage in risky sexual behaviors such as survival sex (exchanging sex for food, money, shelter, drugs, and other needs and wants). Rice et al [43] found that the social connections maintained on Facebook were related to the acceptability of different types of HIV prevention programs. Young and Rice [40] find that using online social networks for partner seeking is associated with an increased risk of sexual behavior among YEH. This trend is especially concerning for YEH, as research suggests a significant portion (25%) of youth engaging in survival sex utilize apps to find partners [44]. The same study found that exchange sex was associated with having recent HIV-positive sex partners and an increased number of multiple sexual partners which has often been found to be associated with sexually transmitted infections and other risk factors [44].

Literature has shown that it is not only the structure of social network connections but the content of such interactions (e.g., email content and conversation topics) that have an impact on the health of YEH [10,42,45]. Barman-Adhikar et al [10] found that YEH used social media to converse about a range of topics. When they talked about topics such as drugs, drinking, or partying, they were more likely to have multiple concurrent sexual partners. Conversely, when they talked about personal goals, plans, and safe sex, they were more likely to engage in protective sexual behaviors. These findings suggest the importance of using social media as a resource for social workers to assess this hard-to-reach group and connect them with care.

The studies reviewed above depend on survey data, which is often flawed and difficult to collect, especially for this group. Furthermore, the social media usage data obtained from the participants are indirect measures obtained solely via self-report questionnaires. Such data depends on the retrospective recall of the respondents and may not reflect the most accurate information. Nevertheless, these studies have established the connection between social media usage and health-related behavior among YEH. To the best of our knowledge, the only work on YEH's drug usage behavior that used data collected directly from social media is conducted by Dou et al [46]. In this study, the authors combined both social media data from Facebook and survey responses from YEH to predict drug usage behavior [47,48]. While they used a combination of social media texts and survey data to build a model with commonly accepted levels of validity, we aim to explore ways of predicting drug usage behavior through social media data solely. We argue that this is a requirement

to study this population, as it is not always feasible to obtain data via other methods like surveys from YEH due to their transient lifestyle. Therefore, we strive to develop a framework that can improve the performance of ML models that only use social media data. We do so by extracting more features than simply the text in the social media posts: the sentiment detected from the posts, the topics expressed in the posts, as well as the reactions from peers on the social media.

In summary, our research objective is to develop a machine-learning-based framework that can identify homeless youth who may be at risk of substance use through their conversations with peers on social media. With this framework, we hope to provide early warnings to social workers and other health professionals to assist them in prioritizing and providing timely interventions to YEH with the most need.

## Methods

## Recruitment and Data

We recruited homeless youth at a non-profit organization located in the Ballpark neighborhood of downtown Denver, a city in the western US with a population of about 800,000 in a metropolitan area of about 2.5 million, between July 2017 to March 2018. Recruiters were present at the agency for over six months, for the duration of service provision hours to approach and screen youth and invite participation. Youth who were interested in the study were screened for eligibility and whether they owned a Facebook (FB) profile for at least a year. For youth who met eligibility criteria, we sought informed consent for participation and obtained their FB account information.

From these participants, we collected two types of data: (1) using our social media crawler, we collected participants' social media conversations for the past year before they were recruited in 2017, including their FB posts as well as the comments and reactions to these posts; (2) we asked the participants to complete a survey on their demographic information, health conditions, sexual behaviors, and substance usage behaviors. Table A1 in Appendix A provides survey question categories with sample questions. Specifically, participants reported whether they have used marijuana, cocaine, coke, crack, heroin, methamphetamine, and/or ecstasy in the last 30 days. Table A2 in Appendix A provides substance usage questions and answer codes. In total, we collected 135,189 FB conversations (including both posts and comments) from these 133 participants. On FB, each post can also receive "reactions", which are extensions of the Link button to allow FB users to share their reactions to a post. These reactions include *Like*, *Love*, *Wow*, *Haha*, *Sad*, and *Angry*. We collected such data as well.

Three of the 133 participants had duplicated FB IDs and were eliminated. Table A3 provides the summary of substance use among these 130 participants. Out of the 130 valid participants who finished the survey, although only 13 (10%) of them indicated in their survey that they did not spend any time on a social media app on a typical day, 46 (35%) of them did not post anything on their FB timeline in the past year. This resulted in 84 participants with their FB posts and comments. We then removed the FB posts without any meaningful textual messages, as well as posts that had missing value in terms of reactions they received. This resulted in 18,788 posts authored by 84 participants. In total, these posts had 19,680 comments and 81,052 reactions. Table 1 provides the summary statistics.

Table 1. Summary statistics

| Data Sources | Characteristics | Mean | Std dev | Min | Max |
|---|---|---|---|---|---|
| **Individual Level Data (cross-sectional)** | | | | | |
| From Survey (84 observations) | Age | 20.58 | 1.94 | 18 | 24 |
| | % Male | 58.3% | | | |
| | % Attending school | 15.5% | | | |
| | % Currently working | 31.0% | | | |
| | Ethnicity | 40.5% white<br>21.4% African American<br>14.3% Hispanic or Latino<br>2.4% American Indian<br>1.2% Asian or Pacific Islander<br>20.2% More than one race or others | | | |
| **Facebook Data (Time stamped)** | | | | | |
| FB Posts (18,788 observations) | # of posts per person | 223.7 | 268.8 | 1 | 1525 |
| FB Comments (19,680 observations) | # of comments per person | 234.2 | 356.1 | 0 | 2415 |
| Reactions (81,052 observations) | # of reactions per person | 964.9 | 1802.5 | 2 | 12155 |
| | 84.1% Like, 8.6% Love, 0.7% Wow, 5.0% Haha, 1.2% Sad, 0.3% Angry | | | | |

## Machine Learning Feature Identification

To identify relevant features for building the ML framework to predict participants' substance use behavior, we drew from literature and theories such as the social sharing of emotions theory and social support theory. We describe each feature set in the following.

### Feature Set 1: Social Media Engagement

The first set of features is the youth's social media engagement behavior, such as the frequency of posting and the length of such posts. Studies have shown an association between social media engagement and risky behavior seeking among adolescents (see [49] for a review). For example, Moreno and Whitehill [50] proposed a Facebook influence model that argued for the positive association between social media use and adolescents' susceptibility to risky behaviors through a peer influence mechanism. The association can also be explained by the displacement hypothesis [51] - social media use can replace time spent on health-related behaviors including in-person social interactions and physical activity. As a result, the increases in social media use among adolescents may have displaced engagement in risky behaviors, such as excess alcohol consumption and illicit drug use [52].

### Feature Set 2: Social Sharing of Emotions in Youth's Posts

The second set of features is the emotions or sentiments in homeless youth's social media posts. Social sharing of emotions refers to the verbal expression of emotion to others [11]. Users of social media share their emotions to seek social support [53], enhance their emotional states [54], and regulate their emotions [55,56].

We include this feature for two reasons. First, the *sharing* behavior itself can be indicative of the ability to prevent or reduce unhealthy behaviors such as substance use. Literature has shown that sharing emotions allows others to provide empathy and support [11]. Such support helps people cope with stress, engage in healthy behaviors [57], and thus reduce stress-coping responses such as substance use [58]. Social support is also shown to be beneficial to YEH – YEH often seek social

support from their peers, at least in the offline setting, which results in positive outcomes such as lower rates of substance use [59].

Second, the *emotion* expressed in social media content is indicative of substance use behavior. Research in substance use has adopted the emotion theory perspective [60] and demonstrated the causal role emotional states play in substance use behavior. For example, studies have documented a high level of emotional instability among substance abusers [61,62], as well as the connection between emotional states and substance use [63–65]. In particular, negative emotion has been observed to be associated with substance use behavior, the inability to withdraw, and the tendency to relapse. For example, studies found that negative mood is associated with a craving for alcohol [63,64]. This association may be explained by the common belief in the ability of substances to alleviate negative moods and reduce stress [66,67]. Negative emotions can also be associated with continuous use and potential relapse [63,65]. Tiffany [65] reported that negative emotional states can interfere with a conscious effort to interrupt automatic drug use behavior, therefore leading to continuous use or relapse.

### Feature Set 3: Topics in Homeless Youth' Posts

The next set of features is the topics mentioned in homeless youth' social media posts. There are two purposes for extracting topics discussed in social media posts. First, we believe that these topics are suitable features for predicting substance use behavior based on findings in related work. Psychoanalytic theory has suggested the relationship between the content of our conversations and our social behaviors [68]. Several studies have found empirical evidence of the relationship between topics mentioned in social media content and users' substance use behavior [32,69]. For homeless youth in particular, Barman-Adhikari et al [70] reported an association between topics discussed by YEH on social media sites and their risky sexual behaviors. For example, talking about drugs, drinking, or partying online is associated with an increased likelihood of engaging in concurrent sex. This study did not examine the relationship between such topics and substance use behavior. Nevertheless, it confirmed certain topics discussed in online social media can be related to offline unhealthy or risky behavior.

Second, these topics can reveal valuable information about the psychological states of the authors of the social media posts – in this case, the YEH. The language one uses in both speech and writing can reveal his or her psychological and social states [71,72]. The users of the framework, such as social workers and researchers, can review these topics and incorporate them in follow-up interviews and surveys to better understand YEH's situations. This framework can be applied to other online communities to provide automatic topic extraction and summarization in social conversations.

### Feature Sets 4 and 5: Social Sharing Interactions with Peers

Social sharing of emotions stimulates social interaction. Such social sharing interactions can strengthen social bonds and end in enhanced social integration [11]. In the case of homeless youth, when they share their emotions and opinions on social media platforms, they also receive social interactions from their peers. Take Facebook as an example, such interactions include quantitative reactions such as *Like* and *Proud*, and qualitative comments to their posts. There is well-established literature that shows the positive impact of social support on health behaviors and health outcomes for the general population [57,73–75]. Studies that use social media conversations to predict other mental risk behaviors also used interaction as a predictive feature [36,37]. In particular, literature has suggested the importance of social support in reducing substance use [12]. For homeless youth, scholars found that homeless adolescents tend to seek emotional support from their peer-based networks [76]. When YEHs share their emotions on social media and receive support through reactions and comments, they may feel cared for and bonded with others, which can reduce their need for substance use. Therefore, we included both *Reactions* and *Comments* YEH received from their peers as the last two feature sets in our machine learning (ML) framework. Particularly, for each

YEH's posts, we include the number of reactions to his/her posts, the number of comments to these posts, as well as the average sentiment of all these comments. Table 2 summarizes the feature sets and the referencing theory of including each feature set.

Table 2. Framework features and guiding theories

| Features | Guiding Theories | Content |
|---|---|---|
| **YEHs' Social Sharing** | | |
| **Social Media Engagement** | Facebook influence model [50], displacement hypothesis [51] of social media | How often a YEH uses the online platform to post content and the average length of his/her posts |
| **Social Sharing of Emotions** | Social sharing of emotions [11], emotion theory [60] | Each YEH receives a score of overall sentiment based on all his/her posts |
| **Social Sharing of Topics** | Psychoanalytic theory [68] | Each YEH is represented by a vector indicating the proportion of each of the topics in all his/her posts |
| **Social Sharing Interaction with Peers** | | |
| **Peers' Reactions to YEH's posts** | Social support theory [12] | Reactions such as Like, Love, Wow to each YEH's posts, depending on the social media sites |
| **Peers' Comments on YEH's posts** | Social support theory [12] | How actively a YEH's posts get attention<br><br>Each YEH receives a score of overall sentiment based on all comments on his/ her posts |

## Feature Extraction

Among these five feature sets, the social media engagement and reaction from peers do not need further extraction. In the following, we focus on how we utilize natural language processing techniques to extract sentiment and topics from the posts, as well as the sentiment from the comments.

### Text Preprocessing

To prepare the texts in our dataset for sentiment analysis and topic analysis, we first need to preprocess the texts. We first removed web links, numbers, and names because such information does not contribute to the understanding of the sentiment and topics of texts. We did not remove punctuations because some punctuation such as exclamation points may carry sentiment weights. We then tokenized the texts into words and phrases and removed common stopwords such as "and", "the", and "a". The output tokens are then ready for sentiment analysis because we want to preserve information such as punctuation, emojis, as well as words in their original forms. For example, "happy", "happiest" and "HAPPY" may have different intensities of emotion. We then performed further text preprocessing for topic modeling. First, we removed all the punctuation. Then we converted all the texts into lowercase. Then we performed lemmatization by reducing each word variant to its base form. Lemmatization is preferred over stemming because it could produce more readable words, as easy-to-interpret output is desirable in topic modeling. Next, we performed part-of-speech (POS) tagging for words and kept the following POS tags: nouns, verbs, adjectives, and adverbs. Finally, we believe it is important to prune candidate words to reduce noise and vocabulary size because terms with high frequency and low frequency are not very useful in topic modeling

[77]. We pruned unigrams and bigrams with high and low frequencies (for example, the words that occurred in more than 70% or less than 1% of the documents) for topic modeling to reduce noise and vocabulary size. We experimented with different frequency cutoffs (between 0% and 100%, between 1% and 70%, and between 5% and 80%) in the topic modeling step and compared the results, which will be discussed later.

### Sentiment Analysis

We used sentiment analysis (SA) to detect the emotions and sentiments in FB texts [78]. We used VADER, a lexicon and rule-based sentiment analysis tool to perform sentiment analysis on our dataset of FB posts. To calculate the sentiment intensity expressed in each post, we first identified words in the conversation that have a sentiment orientation by using VADER's sentiment lexicon [79]. This lexicon comprises lexical features such as words, punctuation, phrases, and emoticons, each assigned with a valence score [80]. A valence score describes the degree of sentiment intensity, from the most negative (-1) to most positive (+1). Then we computed an overall sentiment score for a post by summing the valence scores of all the lexicons (including words, phrases, punctuations, and emojis) detected within the text, adjusted according to grammatical and syntactical rules such as negation and degree intensifiers. These intensifiers are called booster words, such as "extremely" and "marginally" that impact sentiment intensity by either increasing or decreasing the intensity. Finally, VADER normalized this score between -1 and 1. We described this normalization process in Appendix B.

To evaluate the performance of VADER, we randomly selected 300 messages and manually categorized each one into three sentiment categories: positive, negative, and neutral. We then labeled the sentiment of each message predicted by VADER using the following rules: a VADER score of 0 indicated neutral sentiment; a VADER score between 0 and 1 indicated positive sentiment; and a VADER score between -1 and 0 indicated negative sentiment. Finally, we compared the human-annotated sentiment categories with the VADER-predicted sentiment categories and found an agreement rate of 70%. Table A4 provides the performance of VADER classification for all messages, as well as positive, negative, and neutral messages separately.

We then ran SA to the 18,788 FB posts authored by our survey participants, as well as 19,680 comments made to these posts. After we calculated the sentiment scores of all FB posts by our participants, we aggregated such sentiment values to the individual level. Each participant is represented by a score of overall sentiment based on all her posts, and a score of overall sentiment based on all the comments made to her posts.

## Topic Analysis

To address the challenge of topic-modeling short texts on social media sites, we employed the author-topic model [81], which extends the Latent Dirichlet Allocation (LDA) method [82]. It can be viewed as aggregating messages for a user before topic modeling [83]. We did this for two reasons. First, LDA assumes each document is a mixture of various topics while a single social media post (such as a Facebook timeline update) usually only contains a single topic. Combining the posts of an author into one document allows the co-occurrence of multiple topics. Second, the author-topic model allows for the modeling of user interest, which suits our purpose of modeling each participant. Empirically, studies have demonstrated the superior performance of topic models learned from aggregated messages by the same user in short-text environments [84].

Because LDA does not pre-define the number of topics, we needed to determine the best number of LDA topics for our dataset. We varied the number of LDA topics from 5 to 25. We used three commonly used criteria for selecting the optimized number of LDA-generated topics: the coherence score of topics, the rate of perplexity change, and the interpretability of topics [85,86]. First, for each number of topics, we calculated the average coherence scores of all the topics [87]. A topic is coherent if all or most of the words are related. A high average coherence score indicates better topic quality. Therefore, the number of topics corresponding to a *high* average coherence score is a good candidate for the optimized number of topics. Second, we calculated the rate of perplexity change [83]. The rate of perplexity change (RPC) for topic number $t_i$ is calculated as in equation (1):

$$\text{RPC(i)} = \frac{P_i - P_{i-1}}{t_i - t_{i-1}} \quad (1)$$

Where $P_i$ is the perplexity score [88] when the LDA model generates $t_i$ topics. According to Zhao et al [85], the number of topics corresponding to the change of slope for RPC versus the number of topics is considered a good candidate for the optimized number of topics. That is, we should look for "elbows" where the $RPC_i$ is smaller than $RPC_{i+1}$. Therefore, the number of topics corresponding to a *low* RPC is a good candidate for the optimized number of topics. Third, we reviewed the top five representative words of each topic and interpreted them by experience [83].

Figure A1 in Appendix A plots both the coherence score and perplexity change rate vs. the number of topics. According to Figure A1, 9 and 19 are good candidates for the best number of topics with a high coherence score and a low RPC. We reviewed the top words and interpreted them for the resulting 9 topics and 19 topics. Based on our review, we chose 19 topics as the best number of topics because the latter had a higher coherence score and gave us more interpretable representative words and underlying topics. This number was comparable to the number of topics with similar research in FB status updates (e.g., 25 topics in [77]). We also compared the performance of different sets of word candidates—unigrams and bigrams—as inputs for topic modeling. These sets included unigram and bigrams that occurred in all documents, those that appeared in more than 1% but less than 70% of the documents, and those that appeared in more than 5% but less than 80% of the documents. The results were

similar in terms of determining the optimal number of topics.

Once we determined the ideal number of topics, we represented each participant as a vector of topics, using the proportion of different topics in all her posts.

### Text Vectorization

Finally, we developed a vector representation of each participant by vectorizing their posts. To do this, we first combined all the posts by the same participant into one single document. On average, each participant had 224 posts. Word embeddings or encodings are commonly used in studies that use social media data to predict health-related behaviors [89]. We utilized the Global Vectors for Word Representation embeddings (Glove), which provide pre-trained word vectors [90]. These word vector representations were obtained by aggregating global word-word co-occurrence statistics that show how frequently a word appears in a context. This model is commonly used with social media texts [91], due to its pre-trained nature, adaptability to domain-specific corpora, and its ability of handling sparse data, which is prevalent in the domain of social media texts. Each message was converted into a sequence of word vectors using Glove. To handle variable-length sequences, all sequences were padded to the same length: the maximum length of all sequences, by adding zeros. We will combine this vector with the other five feature sets identified above.

## Machine Learning Substance Use Behavior

Using the method described above, we represented each participant as a vector of the abovementioned features: the frequency of one's posting, the average length of one's posts, the average sentiment of one's posts, the proportion of different topics in one's posts, the average reactions one received, and the average sentiment in all the comments one received. Finally, we also included one more feature: the word embeddings of one's posts. We then joined this data with participants' survey responses so that we could use these features to predict participants' substance use behavior.

Among the 84 YEHs in our final dataset, 58 of them (69% of YEHs) were marijuana users. Table A5 provides the summary of substance use distribution among these 84 final participants. Note that the percentage of users for the same drug is not always consistent between the original participant pool of 130 participants and the final participant pool of 84 participants who had an active FB timeline in the past year. A follow-up study can investigate the discrepancy between the two groups of YEH (who had an active FB timeline in the prior year vs. those who did not have an active FB timeline in the prior year) in terms of their substance use patterns. Table A8 provides the distribution of substance use by gender group and age group.

To ensure robust performance evaluation and avoid overfitting, we employed a stratified k-fold cross-validation method [92,93]. We set the number of folds to 3 due to the relatively small sample size. When splitting the data, the class distribution in the training and test sets is set to be the same as in the full dataset. The random seed for shuffling was set to a predefined value to ensure the reproducibility of our results.

We performed the machine learning prediction in two steps. Because we wanted to efficiently combine the word vector features with other numeric feature sets without losing the contextual information of the word vectors, we first employed a neural network using word embeddings from the posts solely as the input to predict marijuana use. The output from this first model was then used as a prediction feature and combined with the other numeric feature sets. Subsequently,

we applied all these features together to a variety of machine learning models. We chose this approach because it allows us to leverage the strengths of both neural networks and traditional machine learning models. We describe each step in detail below.

First, we used TensorFlow and Keras to construct a neural network (NN) model. This model first accepted pre-trained GloVe word embeddings of a YEH's posts as the input. We employed a bidirectional long short-term memory (LSTM) layer to capture both forward and backward sequential dependencies in the text. The LSTM output was passed through a dense layer with Rectified Linear Unit (ReLU) activation, and the output was then flattened before making the final prediction using a sigmoid-activated dense layer.

Second, we used the output of the neural network model as one prediction feature and combined it with the other numeric feature sets: social media engagement, social sharing of emotions, post topics, reactions, and comments. These combined features were then applied to a variety of machine-learning models. One challenge in predicting the substance behavior of YEH is that we often only have access to very small datasets due to the transient nature of this group. Based on the nature of the data, we used *bagging* [94] and *ensemble learning* [95] so that the users could draw bootstrap samples from the data and perform the same estimator for each sample. The overall prediction can be obtained by simple voting. This can reduce the variance and stabilize the performance of classifiers when working with small training datasets [96]. We drew 1000 bootstrap samples from the data and performed the same estimator for each sample. The overall prediction was obtained by simple voting. We employ bagging to three base classifiers: decision tree, logistic regression, and support vector classifier (SVC). Decision tree and SVC are suitable because they are both popular models for text classification on social media [97].

## Results

### Feature Extraction Results

In the last section, we discussed the natural language processing method of extracting sentiment and topics from YEH's posts and comments. Figure 1 shows the distribution of the average sentiment scores of posts for the participants. Figure 2 shows the distribution of the average sentiment scores of posts for marijuana users and non-users. Similarly, Figure 3 shows the distribution of the average sentiment scores of comments for the participants. Figure 4 shows the distribution of the average sentiment scores of comments for marijuana users and non-users. Based on the figures, non-users have a high proportion of posts and comments with positive sentiment scores, compared to users.

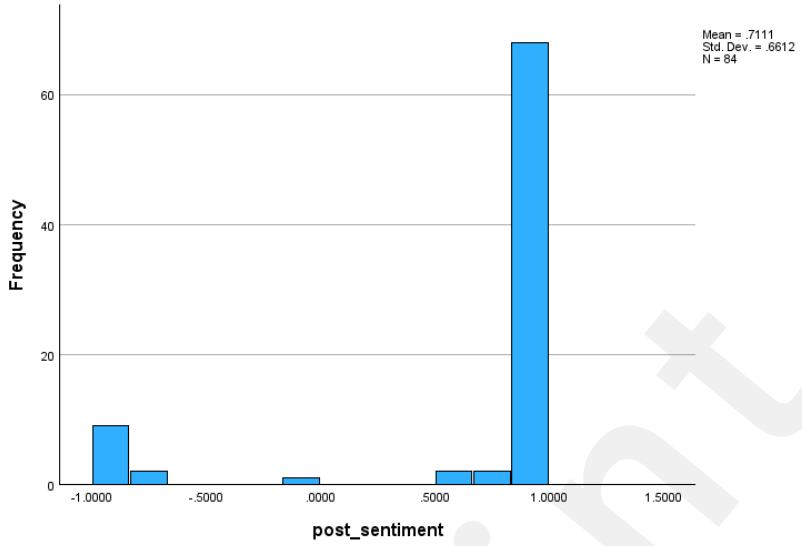
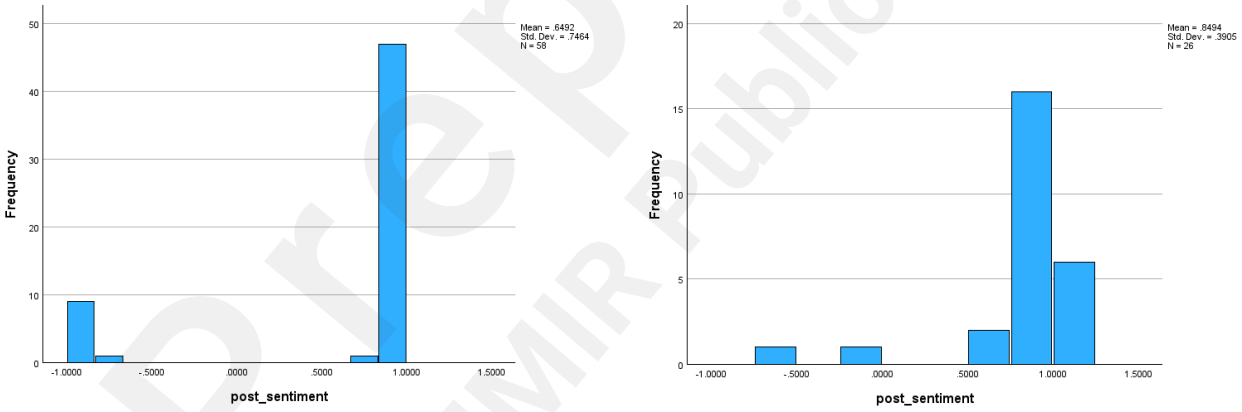



**Figure 1. Distribution of average sentiment scores of all participants' posts**




**Marijuana Users Posts** **Non-users Posts**

**Figure 2. Distribution of average sentiment scores of all participants' posts marijuana users vs. non-users**

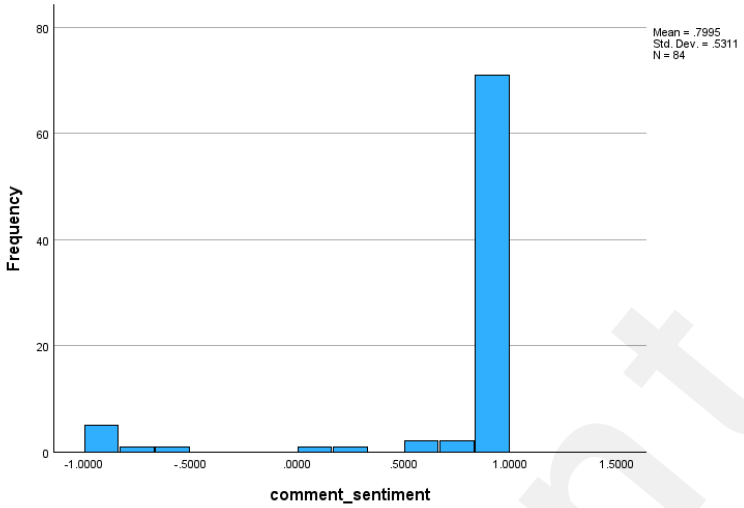
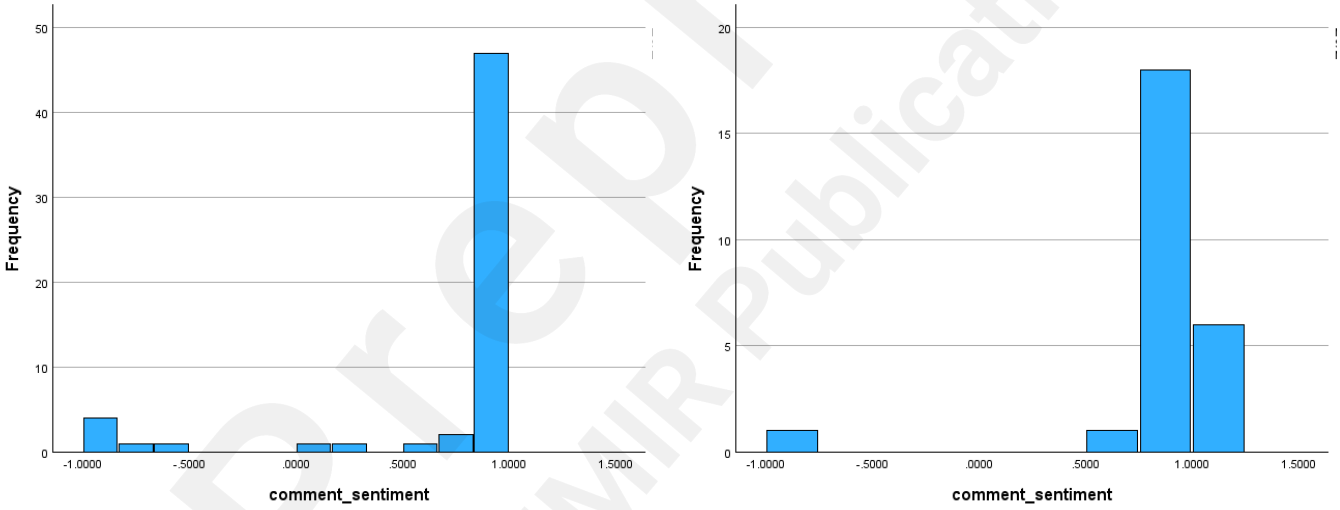



**Figure 3. Distribution of average sentiment scores of all participants' comments**




**Marijuana Users Posts** **Non-users Posts**

**Figure 4. Distribution of average sentiment scores of all participants' comments marijuana users vs. non-users**

For topic modeling, we picked 19 as the ideal number of topics, based on coherence score, perplexity score, and interpretability. Table A6 provides the top five topics (ranked by frequency in the documents), their top representative words, the latent topic themes based on our interpretation, and the frequency of these topics mentioned by marijuana users and non-users. It is interesting to note that three of the five topics – work, swear, and female-related, have been proven to be related to substance consumption among the general population [32,69].

## Machine Learning Results

Table 4 shows that the bagged decision tree provides the most accurate prediction compared to others and is our model choice going forward. We compared the performance of our model with a benchmark model developed by Tabar et al. [98], who used survey data (such as demographic information and criminal history) to predict substance use in a similar population. In addition, we applied each of the three models (bagged decision tree, bagged SVC, and bagged logistic regression) using the survey data in our dataset. The features used in the survey data and the performance are reported in Table A7 in Appendix A. Our results showed that our feature sets outperformed the feature set of survey information when the same model is applied (with the exception of bagged SVC when the AUC is 0.50 for both feature sets). Our result showed that our framework can use social media data to predict certain substance use with better performance.

Table 4. Performance of different ML models for predicting marijuana use

| Model | AUC | Accuracy | Precision | Recall | F1 |
|---|---|---|---|---|---|
| **NN+ Bagged Decision Tree** | **0.72** | **0.81** | **0.81** | **0.95** | **0.87** |
| NN+ Bagged SVC | 0.50 | 0.69 | 0.69 | 1.00 | 0.82 |
| NN+ Bagged Logit | 0.66 | 0.76 | 0.77 | 0.93 | 0.84 |
| Benchmark model by [98] | 0.72 | 0.69 | 0.73 | 0.79 | 0.76 |

Testing size: 28

## Ensuring Fairness

Given that substance use can be disproportionally associated with certain populations, it is important to mitigate potential gender, racial, and socioeconomic biases in our framework. We followed the “fairness by design” strategy [99] to ensure fairness throughout the study. First, our team included a social work researcher, who worked to minimize the biases during data collection, model building, and result interpretation. Second, we chose to exclude the survey data from the feature selection process, due to its potential to contain both explicit and implicit demographic biases. However, even with such exclusion, social media contents often contain language, topics of interest, and other contextual information that could potentially reflect a user's demographic characteristics. These implicit cues can inadvertently introduce biases into our model. Therefore, we conducted a post-hoc analysis to evaluate the model’s performance across different demographic groups, ensuring that it performs equitably. Specifically, we evaluated the false positive rate (FPR) for each gender and age segmentation. This is because we were aware of the stigma surrounding YEH and drug use and wanted to identify false flags. Table 5 summarizes the evaluation results by age and gender in the test dataset, and Table A8 in Appendix A summarizes the marijuana use distribution in the original dataset by age and gender. We found that the FPR is slightly higher among female participants as well as participants who

were 21 years old or above. Users of this framework should deploy it cautiously and avoid over-generalization, especially for these two groups.

| | Accuracy | Precision | Recall | False-positive Rate |
|---|---|---|---|---|
| **Group by Gender** | | | | |
| **Male** | 0.85 | 0.87 | 0.93 | 0.10 |
| **Female** | 0.73 | 0.72 | 0.94 | 0.23 |
| **Group by Age** | | | | |
| **Below 21** | 0.88 | 0.85 | 1.00 | 0.12 |
| **21 and above** | 0.74 | 0.78 | 0.89 | 0.18 |

# Discussion

## Principal Findings

Current studies that leverage social media data to predict users' substance behaviors usually neglect users' interactions with their peers in the community, as well as the semantic meanings of their posts, such as the topics expressed. In this paper, we develop a social media-based framework that applies NLP techniques and ML models to predict YEH's substance use behavior using their social media posts and interactions. We built on theories such as social sharing of emotions and social support theory to develop an effective set of features and demonstrated the effectiveness of our framework for practical use in detecting YEH at risk of using marijuana on social media platforms. Our best model reaches an accuracy of 0.81 of an AUC of 0.72 when predicting marijuana use. We have made a few notable findings.

First, we used a combination of social media posts, comments, and reactions to build a framework that can predict substance use as self-reported by the subjects. Guided by theories such as social support theory and social sharing of emotions, we developed a unique set of features from participants' social media conversations that can be indicative of substance use behaviors.

Second, we found that the sentiment of all FB posts that are authored by our survey respondents is overall positive. A similar trend has been observed among housed youth [100,101]. We show that YEH do not necessarily express a more negative sentiment on social media sites than their housed counterparts. Prior studies have shown that sentiment-related indicators from one's social media texts can relate to his/her health characteristics such as mental wellbeing for the more general population [20,69,102–104]. Our framework provides the means to support the observation of the mental well-being of YEH. It is worth noting that although we show a similar sentiment pattern between YEH and other college students, it can be hard to compare the sentiment values reported by different studies due to the different sentiment analysis methods, lexicon, and scales of sentiment values used in each study. Future studies can apply sentiment analysis to FB conversations of both YEH and their housed peers for better comparison.

Third, we also examined topics in YEH's posts extracted by our framework. The most frequent topics are related to relationships, work, swear, females, and lifestyle (see Table 3). We compared this list of observed topics with topics reported in survey responses from our participants. We found a few discrepancies between the two lists. In the survey answers, 32.5% of participants reported talking about drugs, 26.6% reported talking about sex, 26.2% reported talking about

school and/or work, 24% reported talking about family issues, 23.9% reported talking about being homeless, and 5.3% reported talking about goals. While some of these topics are common, especially the focus on work, we were able to reveal a few unique topics of discussion through the use of digital trace data that would not be captured through predetermined survey questions. For example, relationship is one topic that seems to be very important for this group of young people. This likely underscores the methodological and substantive benefits of using social media data. It is interesting to note that several of these topics we found in participants' posts, such as swear and females, are observed to be positively related to substance use among the general population [34,35]. Prior studies have shown that the topics from one's social media texts can relate to his/her health characteristics such as mental well-being. We also compared the distribution of the top five topics between marijuana users and non-users in Table A6 in Appendix A. Overall, marijuana users tended to mention the following four topics more often than non-users: relationship, work, swear, and lifestyle. Non-users, on the other hand, mentioned the female-related topic more often. While these observations highlight interesting trends, it is important to note that the differences in topic distribution may not be statistically significant. Therefore, a follow-up study with more rigorous statistical analysis is recommended to investigate and confirm these discrepancies. Future studies can provide deeper insights into the social and psychological factors associated with marijuana use and help in understanding the underlying reasons for these differences in topic prevalence between users and non-users.

Because our framework provides users the capability of automatic extraction of sentiment and topics from social media conversations, one application of our framework is to examine and compare the sentiment and topics expressed in social media conversations between YEH and their housed peers.

Our research has several notable implications for research in mining social media for substance behavior prediction, and the practice of substance use outreach and prevention.

## Implication for Research

The major contribution of our study is the unique design of a framework that combines a series of theory-guided social-media-based features that can predict YEH's substance use behavior. The proposed feature set achieves the best AUC and accuracy compared to existing methods by prior studies. These results suggest that such theory-guided features can achieve better performance over other models such as word-embeddings and bag-of-words models that do not take the semantic meanings of social media conversations into account. Such results also contribute to the literature on substance use behavior. We found that the sentiment and certain topics in YEH's posts, as well as reactions and comments to these posts, can predict YEH's marijuana use. Researchers can build on this finding to develop instruments for developing a better understanding of YEH's mental states and substance usage tendencies.

We also demonstrate the feasibility of mining digital trace data from social media platforms to predict YEH's health behaviors. We showed that textual interactions among YEH and their friends on social media can serve as a powerful resource to predict their substance usage. We proved that, without survey information, which may introduce gender and racial biases, our ML models can reach an AUC of 0.74 and an accuracy of 0.77 using social media data only.

## Implication for Practice

Substance use disorder is a significant public health concern among YEH which can potentially lead to other health-related problems (e.g., risky sex behaviors, mental problems, and sexually

transmitted diseases). With the increase in funding to free or subsidized treatment facilities from a local, state, or federal level[105], our work provides a foundational strategy for how those funds could be applied effectively to improve the outreach of these facilities.

After discussions with several providers in the Denver area, we believe our work will provide an effective complementary tool for facilities' outreach efforts. Community Alcohol Drug Rehabilitation and Education (CADRE) provides relapse prevention therapy services and is currently developing an outreach program. In the future, we seek to work with local providers to determine the extent to which our tool can improve their outreach efforts.

Target users of this framework include school counselors, juvenile diversion programs, homeless shelters, and substance use intervention facilities. We identified three ways that relevant groups, after receiving the appropriate permissions from YEHs, could use our framework. First, they can get a better understanding of YEH's current mental state by reviewing the sentiments and topics extracted from YEH's social media pages. Second, this framework can be used as a preventative measure through the identification of the YEH at risk for substance use. Third, this framework can be used for frequent monitoring and enhancing therapies designed to minimize the likelihood of relapse among YEHs.

Studies have shown the effectiveness of web-based intervention programs (e.g., [106]). Our framework provides the foundation for developing online substance intervention programs in social media communities. Healthcare agencies can work with social media companies to incorporate our framework into the platform through which personalized interventions could be developed and distributed. Given YEH's permissions, such interventions could include appointment reminders, status monitoring services, free education, and the offering of third-party professional assistance.

## Ethical Considerations

Our research acknowledges the inherent privacy risks associated with social media data, which can be highly personal and sensitive [107,108]. To mitigate these concerns, we prioritized ethical data collection practices throughout the study. A cornerstone of our approach was transparency. We obtained informed consent from participants, ensuring they fully understood the data being collected from their social media accounts and who would have access to it. This transparent approach empowers participants to make informed decisions about their data.

Following data collection, all identifiable information was meticulously removed to prevent re-identification. We employed techniques like assigning Personal Identification Numbers to anonymize data during storage and analysis, similar to traditional research practices [109,110]. Furthermore, participants retained the right to withdraw their data at any point. We facilitated this by providing clear contact information and an online option for quick responses to withdrawal requests. Participants could also leverage their social media privacy settings to further restrict data access after collection.

Our research protocol strictly adhered to established ethical guidelines, including those outlined by Institutional Review Boards for data mining on platforms like Facebook. These guidelines emphasize minimizing risks to participants, ensuring a fair subject selection process, and prioritizing data protection throughout the research lifecycle.

We also must ask how this research would be used. That is, how would individuals react if they were knowingly called out by social workers for potential substance abuse even though they did not admit to it, especially if they discovered that it was due to their social media posts? There is a probability that individuals would stop seeking services from the institutions that can help them,

and they might direct other youth experiencing homelessness (YEH) to not seek help from those social services. Other YEH may alter their social media postings or move to other “dark web” social media that are not as easily trackable. Moreover, all predictive models are subject to errors. False positives would be the bigger concern here, as YEH are often ostracized for being “junkies” even if they do not abuse substances. According to [111], predictive algorithms in social services have a history of disproportionately affecting marginalized groups, often exacerbating the very issues they aim to solve.

The ethical implications of using social media surveillance for identifying at-risk youth are significant. As Marwick et al. discussed [112], the act of surveillance can change individuals’ behavior, often leading to increased privacy measures and a decrease in trust towards institutions conducting the surveillance. YEH who are aware that their social media is being monitored might feel their privacy is invaded, which can lead to a range of negative outcomes, including psychological distress and reluctance to engage with supportive services [113]. Furthermore, as argued by Van Dijck [114], the practice of datafication—turning social behaviors into quantifiable data—can dehumanize individuals and overlook the contextual nuances of their actions, leading to misinterpretation and harm.

Given these concerns, it is crucial that any implementation of predictive models in social services includes robust ethical guidelines and continuous oversight. As discussed above, transparency in how data is collected, used, and shared, as well as clear communication with the affected individuals about these processes, can help mitigate some of the adverse effects. For example, YEH who consented to participate should be notified of the means and frequency of their social media monitoring and the subsequent intervention and be able to choose the level and frequency of monitoring and intervention they prefer. They should also be able to terminate their participation at any time. Ensuring that social workers and other practitioners are trained in the ethical use of these tools and are sensitive to potential harm can further reduce the risk of negative outcomes.

## Limitations

As with all research, there are limitations involved and discussion as to how the research can and should be used. The first limitation of this research is the relatively small sample size compared to most social science research. Some of this was due to almost half of the subjects recruited having to be eliminated after the data sets were cleaned. While this number is considered low in lab research, in field research with real at-risk subjects we were happy to get the size that we did. While Denver is by no means a small city and unfortunately does not have a small homeless population, it is possible that other cities with larger homeless populations might yield larger numbers of participants. We believe that the model we created would only continue to have higher levels of accuracy with more subjects.

Another limitation of this study is that the data were collected between 2016 and 2017, preceding the current date. We recognize that social media platforms and communication styles can evolve rapidly [115]. However, the core issues surrounding youth homelessness and their use of online platforms for connection and looking for instrumental needs likely remain relevant as evidenced by studies conducted across different time periods [39,116–118]. This temporal gap may affect the generalizability of our findings to the present day, but existing research suggests that the fundamental needs and behaviors of homeless youth in online communication may not change as swiftly as the platforms themselves [117,118]. This consideration is critical in contextualizing our results within the evolving digital landscape.

While studying homeless youth presents difficulties due to their transient nature and distrust of outsiders [119], this data provides valuable insights. Social media analysis offers a rare window into the online behavior of this often-overlooked population. However, to ensure the continued relevance of our findings, further research is necessary to explore potential shifts in communication patterns among homeless youth on social media platforms.

## Future Research

While it is hard for experts and health professionals to assess YEH's health status due to their transient lifestyle, we provide a framework that can automatically detect sentiment and opinion from social media, which can be subsequently reviewed and analyzed by experts from different research backgrounds. Future research can extend this framework to analyze and predict a variety of health-related behaviors of YEH or analyze the behaviors of a more general population.

This study demonstrates the feasibility of mining digital trace data from social media platforms to predict YEH's health behaviors. Future research can look at other social media platforms such as Instagram and TikTok and investigate other forms of digital trace data such as images and videos.

## Conclusion

*"Things in my life are good and finally getting better things have been rough for me but i know i can get through the hardest of times … but i would like to thank all who have been their for me and helped me through things im coming to a new beginning and would like to still have you their with me and to renew any relations i messed up because of my drug additions i had. im now sober and well i feel great now."*

- Social media posting by another YEH in Denver

The scourge of substance use is a major barrier to moving YEH back into stable living situations. We hope that our framework will be one more tool that social service workers can use to identify those experiencing these hardships and help them get the care they need.

## Acknowledgments

This research is supported by the University of Denver's Professional Research Opportunities for Faculty grant. The authors have no other competing interests to declare that are relevant to the content of this article.

## Abbreviations

AUC: area under the curve
FB: Facebook
FPR: false positive rate
LDA: Latent Dirichlet Allocation
LSTM: long short-term memory
ML: machine learning
NLP: natural language processing
NN: neural network
ReLU: rectified linear unit
SA: sentiment analysis
YEH: youth experiencing homelessness

## References


1. National Institute on Drug Abuse Statistics, 2019. Accessed May 31, 2024. https://nida.nih.gov/research-topics/trends-statistics#supplemental-references-for-economic-costs
2. Baer JS, Ginzler JA, Peterson PL. DSM-IV alcohol and substance abuse and dependence in homeless youth. *J Stud Alcohol*. 2003;64(1):5-14. doi:10.15288/jsa.2003.64.5
3. Nyamathi A, Hudson A, Greengold B, et al. Correlates of substance use severity among homeless youth. *Journal of Child and Adolescent Psychiatric Nursing*. 2010;23(4):214-222.
4. Black RA, Serowik KL, Rosen MI. Associations between impulsivity and high risk sexual behaviors in dually diagnosed outpatients. *Am J Drug Alcohol Abuse*. 2009;35(5):325-328.
5. Baer JS, Peterson PL, Wells EA. Rationale and design of a brief substance use intervention for homeless adolescents. *Addiction Research & Theory*. 2004;12(4):317-334.
6. Brooks RA, Milburn NG, Rotheram-Borus MJ, Witkin A. The system-of-care for homeless youth: Perceptions of service providers. *Eval Program Plann*. 2004;27(4):443-451.
7. Dang MT. We need to pay attention to substance use among homeless youth. *J Addict Nurs*. 2012;23(3):149-151.
8. Hudson AL, Nyamathi A, Greengold B, et al. Health-seeking challenges among homeless youth. *Nurs Res*. 2010;59(3):212-218. doi:10.1097/NNR.0b013e3181d1a8a9
9. Curtis BL, Ashford RD, Magnuson KI, Ryan-Pettes SR. Comparison of smartphone ownership, social media use, and willingness to use digital interventions between generation Z and millennials in the treatment of substance use: cross-sectional questionnaire study. *J Med Internet Res*. 2019;21(4):e13050.
10. Barman-Adhikari A, Rice E, Bender K, Lengnick-Hall R, Yoshioka-Maxwell A, Rhoades H. Social networking technology use and engagement in HIV-related risk and protective behaviors among homeless youth. *J Health Commun*. 2016;21(7):809-817. doi:10.1080/10810730.2016.1177139
11. Rimé B. Emotion elicits the social sharing of emotion: Theory and empirical review. *Emotion review*. 2009;1(1):60-85.
12. Lin N, Dean A, Ensel W. *Social Support, Life Events, and Depression*. Academic Press; 1986.
13. Ovalle A, Goldstein O, Kachuee M, et al. Leveraging social media activity and machine learning for HIV and substance abuse risk assessment: development and validation study. *J Med Internet Res*. 2021;23(4):e22042.
14. Santa Maria D, Padhye N, Yang Y, et al. Drug use patterns and predictors among homeless youth: Results of an ecological momentary assessment. *Am J Drug Alcohol Abuse*. 2018;44(5):551-560.
15. Hammond CJ, Chaney A, Hendrickson B, Sharma P. Cannabis use among US

adolescents in the era of marijuana legalization: a review of changing use patterns, comorbidity, and health correlates. *International review of psychiatry*. 2020;32(3):221-234.
16. Memedovich KA, Dowsett LE, Spackman E, Noseworthy T, Clement F. The adverse health effects and harms related to marijuana use: an overview review. *Canadian Medical Association Open Access Journal*. 2018;6(3):E339–E346.
17. Grajales III FJ, Sheps S, Ho K, Novak-Lauscher H, Eysenbach G. Social media: a review and tutorial of applications in medicine and health care. *J Med Internet Res*. 2014;16(2):e2912.
18. Eysenbach G, others. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. *J Med Internet Res*. 2009;11(1):e1157.
19. Singh T, Roberts K, Cohen T, et al. Social Media as a Research Tool (SMaaRT) for risky behavior analytics: methodological review. *JMIR Public Health Surveill*. 2020;6(4):e21660.
20. Guntuku SC, Yaden DB, Kern ML, Ungar LH, Eichstaedt JC. Detecting depression and mental illness on social media: An integrative review. *Curr Opin Behav Sci*. 2017;18:43-49. doi:10.1016/j.cobeha.2017.07.005
21. Settanni M, Marengo D. Sharing feelings online: Studying emotional well-being via automated text analysis of Facebook posts. *Front Psychol*. 2015;6:1045. doi:10.3389/fpsyg.2015.01045
22. Chau M, Li TMH, Wong PWC, Xu JJ, Yip PSF, Chen H. Finding people with emotional distress in online social media: A design combining machine learning and rule-based classification. *MIS Quarterly*. 2020;44(2):933-955.
23. Yang X, McEwen R, Ong LR, Zihayat M. A big data analytics framework for detecting user-level depression from social networks. *Int J Inf Manage*. 2020;54:102141.
24. Liu D, Feng XL, Ahmed F, Shahid M, Guo J, others. Detecting and measuring depression on social media using a machine learning approach: systematic review. *JMIR Ment Health*. 2022;9(3):e27244.
25. Huang YP, Goh T, Liew CL. Hunting suicide notes in web 2.0-preliminary findings. In: *Ninth IEEE International Symposium on Multimedia Workshops (ISMW 2007)*. ; 2007:517-521.
26. Kumar R, Mukherjee S, Choi TM, Dhamotharan L. Mining voices from self-expressed messages on social-media: Diagnostics of mental distress during COVID-19. *Decis Support Syst*. 2022;162:113792. doi:10.1016/J.DSS.2022.113792
27. Hassanpour S, Tomita N, DeLise T, Crosier B, Marsch LA. Identifying substance use risk based on deep neural networks and Instagram social media data. *Neuropsychopharmacology*. 2019;44(3):487-494.
28. Kosinski M, Stillwell D, Graepel T. Private traits and attributes are predictable from digital records of human behavior. *Proc Natl Acad Sci U S A*. 2013;110(15):5802-5805. doi:10.1073/pnas.1218772110
29. Desrosiers A, Vine V, Kershaw T. “RU Mad?”: Computerized text analysis of

affect in social media relates to stress and substance use among ethnic minority emerging adult males. *Anxiety Stress Coping*. 2019;32(1):109-123.
30. Owen D, Antypas D, Hassoulas A, et al. Enabling early health care intervention by detecting depression in users of web-based forums using language models: Longitudinal analysis and evaluation. *JMIR AI*. 2023;2(1):e41205.
31. Tsugawa S, Kikuchi Y, Kishino F, Nakajima K, Itoh Y, Ohsaki H. Recognizing depression from twitter activity. In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ; 2015:3187-3196.
32. Ding T, Bickel WK, Pan S. Social media-based substance use prediction. *arXiv preprint arXiv:170505633*. Published online 2017.
33. Curtis B, Giorgi S, Buffone AEK, et al. Can Twitter be used to predict county excessive alcohol consumption rates? *PLoS One*. 2018;13(4):e0194290.
34. Marengo D, Azucar D, Giannotta F, Basile V, Settanni M. Exploring the association between problem drinking and language use on Facebook in young adults. *Heliyon*. 2019;5(10):e02523. doi:10.1016/j.heliyon.2019.e02523
35. Ding T, Bickel WK, Pan S. Social media-based substance use prediction. *arXiv preprint arXiv:170505633*. Published online 2017.
36. Shen G, Jia J, Nie L, et al. Depression detection via harvesting social media: A multimodal dictionary learning solution. In: *IJCAI*. ; 2017:3838-3844.
37. Wang X, Zhang C, Ji Y, Sun L, Wu L, Bao Z. A depression detection model based on sentiment analysis in micro-blog social network. In: *Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2013 International Workshops: DMApps, DANTH, QIMIE, BDM, CDA, CloudSD, Gold Coast, QLD, Australia, April 14-17, 2013, Revised Selected Papers 17*. ; 2013:201-213.
38. Rice E, Barman-Adhikari A. Internet and social media use as a resource among homeless youth. *Journal of Computer-Mediated Communication*. 2014;19(2):232-247. doi:10.1111/jcc4.12038
39. VonHoltz LAH, Frasso R, Golinkoff JM, Lozano AJ, Hanlon A, Dowshen N. Internet and social media access among youth experiencing homelessness: mixed-methods study. *J Med Internet Res*. 2018;20(5):e184.
40. Young SD, Rice E. Online social networking technologies, HIV knowledge, and sexual risk and testing behaviors among homeless youth. *AIDS Behav*. 2011;15(2):253-260. doi:10.1007/s10461-010-9810-0
41. Rice E, Monro W, Barman-Adhikari A, Young SD. Internet use, social networking, and HIV/AIDS risk for homeless adolescents. *Journal of Adolescent Health*. 2010;47(6):610-613. doi:10.1016/j.jadohealth.2010.04.016
42. Barman-Adhikari A, Rice E. Sexual health information seeking online among runaway and homeless youth. *J Soc Social Work Res*. 2011;2(2):89-103. doi:10.5243/jsswr.2011.5
43. Rice E, Tulbert E, Cederbaum J, Barman Adhikari A, Milburn NG. Mobilizing homeless youth for HIV prevention: A social network analysis of the acceptability of a face-to-face and online social networking intervention.

*Health Educ Res*. 2012;27(2):226-236. doi:10.1093/her/cyr113
44. Srivastava A, Rusow JA, Holguin M, et al. Exchange and survival sex, dating apps, gender identity, and sexual orientation among homeless youth in Los Angeles. *J Prim Prev*. 2019;40:561-568.
45. Calvo F, Carbonell X, others. Using Facebook for improving the psychological well-being of individuals experiencing homelessness: experimental and longitudinal study. *JMIR Ment Health*. 2018;5(4):e9814.
46. Dou Z, Barman-Adhikari A, Fang F, Yadav A. Harnessing social media to identify homeless youth at-risk of substance use. In: *35th AAAI Conference on Artificial Intelligence*. ; 2021.
47. Whitaker C, Stevelink S, Fear N. The use of Facebook in recruiting participants for health research purposes: a systematic review. *J Med Internet Res*. 2017;19(8):e290.
48. Capurro D, Cole K, Echavarr\'\ia MI, et al. The use of social networking sites for public health practice and research: a systematic review. *J Med Internet Res*. 2014;16(3):e2679.
49. Vannucci A, Simpson EG, Gagnon S, Ohannessian CM. Social media use and risky behaviors in adolescents: A meta-analysis. *J Adolesc*. 2020;79:258-274.
50. Moreno MA, Whitehill JM. Influence of social media on alcohol use in adolescents and young adults. *Alcohol Res*. 2014;36(1):91.
51. Kraut R, Patterson M, Lundmark V, Kiesler S, Mukophadhyay T, Scherlis W. Internet paradox: A social technology that reduces social involvement and psychological well-being? *American psychologist*. 1998;53(9):1017.
52. Lewycka S, Clark T, Peiris-John R, et al. Downwards trends in adolescent risk-taking behaviours in New Zealand: Exploring driving forces for change. *J Paediatr Child Health*. 2018;54(6):602-608.
53. Burke M, Develin M. Once more with feeling: Supportive responses to social sharing on Facebook. In: *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ; 2016:1462-1474.
54. Myrick JG. Emotion regulation, procrastination, and watching cat videos online: Who watches Internet cats, why, and to what effect? *Comput Human Behav*. 2015;52:168-176.
55. Bazarova NN, Choi YH, Schwanda Sosik V, Cosley D, Whitlock J. Social sharing of emotions on Facebook: Channel differences, satisfaction, and replies. In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ; 2015:154-164.
56. Vermeulen A, Vandebosch H, Heirman W. # Smiling,# venting, or both? Adolescents' social sharing of emotions on social media. *Comput Human Behav*. 2018;84:211-219.
57. Schwarzer R, Leppin A. Social support and health: A theoretical and empirical overview. *J Soc Pers Relat*. 1991;8(1):99-127.
58. Cohen S. Social relationships and health. *American psychologist*. 2004;59(8):676.
59. Barman-Adhikari A, Bowen E, Bender K, Brown S, Rice E. A social capital approach to identifying correlates of perceived social support among

homeless youth. *Child Youth Care Forum*. 2016;45(5):691-708.
60. Quirk SW. Emotion concepts in models of substance abuse. *Drug Alcohol Rev*. 2001;20(1):95-104.
61. McCormick RA, Dowd ET, Quirk S, Zegarra JH. The relationship of NEO-PI performance to coping styles, patterns of use, and triggers for use among substance abusers. *Addictive Behaviors*. 1998;23(4):497-507.
62. Barnes GE. Clinical and prealcoholic personality characteristics. *The biology of alcoholism*. Published online 1983:113-195.
63. Cooney NL, Litt MD, Morse PA, Bauer LO, Gaupp L. Alcohol cue reactivity, negative-mood reactivity, and relapse in treated alcoholic men. *J Abnorm Psychol*. 1997;106(2):243.
64. Childress AR, Ehrman R, McLellan AT, MacRae J, Natale M, O'Brien CP. Can induced moods trigger drug-related responses in opiate abuse patients? *J Subst Abuse Treat*. 1994;11(1):17-23.
65. Tiffany ST. A cognitive model of drug urges and drug-use behavior: role of automatic and nonautomatic processes. *Psychol Rev*. 1990;97(2):147.
66. Cooper ML, Frone MR, Russell M, Mudar P. Drinking to regulate positive and negative emotions: A motivational model of alcohol use. *J Pers Soc Psychol*. 1995;69(5):990.
67. Cooper ML, Russell M, Skinner JB, Frone MR, Mudar P. Stress and alcohol use: moderating effects of gender, coping, and alcohol expectancies. *J Abnorm Psychol*. 1992;101(1):139.
68. Rapaport D. The structure of psychoanalytic theory. *Psychol Issues*. 1960;2(2):1–158.
69. Marengo D, Azucar D, Giannotta F, Basile V, Settanni M. Exploring the association between problem drinking and language use on Facebook in young adults. *Heliyon*. 2019;5(10):e02523. doi:10.1016/j.heliyon.2019.e02523
70. Barman-Adhikari A, Rice E, Bender K, Lengnick-Hall R, Yoshioka-Maxwell A, Rhoades H. Social networking technology use and engagement in HIV-related risk and protective behaviors among homeless youth. *J Health Commun*. 2016;21(7):809-817. doi:10.1080/10810730.2016.1177139
71. Pennebaker JW, Mehl MR, Niederhoffer KG. Psychological aspects of natural language use: Our words, our selves. *Annu Rev Psychol*. 2003;54(1):547-577.
72. Pennebaker JW. The secret life of pronouns. *New Sci (1956)*. 2011;211(2828):42-45.
73. Chiu CM, Huang HY, Cheng HL, Sun PC. Understanding online community citizenship behaviors through social support and social identity. *Int J Inf Manage*. 2015;35(4):504-519.
74. Bloom JR. The relationship of social support and health. *Soc Sci Med*. 1990;30(5):635-637.
75. Cohen SE, Syme SI. *Social Support and Health.* Academic Press; 1985.
76. Whitbeck LB, Hoyt DR. *Nowhere to Grow: Homeless and Runaway Adolescents and Their Families*. Routledge; 2017.
77. Wang YC, Burke M, Kraut R. Gender,topic, and audience response: An analysis of user-generated content on Facebook. In: *Conference on Human*

*Factors in Computing Systems - Proceedings*. ; 2013:31-34. doi:10.1145/2470654.2470659
78. Pang B, Lee L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: *Proceedings of the ACL*. ; 2004.
79. Vader Sentiment Analysis lexicon. Accessed May 28, 2024. https://github.com/cjhutto/vaderSentiment/tree/master/vaderSentiment
80. Hutto CJ, Gilbert E. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In: *Proceedings of the International AAAI Conference on Web and Social Media*. ; 2016:8(1), 216-225.
81. Rosen-Zvi M, Griffiths T, Steyvers M, Smyth P. The author-topic model for authors and documents. *arXiv preprint arXiv:12074169*. Published online 2012.
82. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *Journal of Machine Learning Research*. 2003;3(1):993-1022. doi:10.1162/jmlr.2003.3.4-5.993
83. Xu Z, Ru L, Xiang L, Yang Q. Discovering user interest on twitter with a modified author-topic model. In: *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*. Vol 1. ; 2011:422-429.
84. Hong L, Davison BD. Empirical study of topic modeling in twitter. In: *Proceedings of the First Workshop on Social Media Analytics*. ; 2010:80-88.
85. Zhao W, Chen JJ, Perkins R, et al. A heuristic approach to determine an appropriate number of topics in topic modeling. In: *BMC Bioinformatics*. Vol 16. ; 2015:1-10.
86. Deng Y, Zheng J, Khern-am-nuai W, Kannan K. More than the quantity: The value of editorial reviews for a user-generated content platform. *Manage Sci*. 2021;68(9):6865-6888.
87. Aletras N, Stevenson M. Evaluating topic coherence using distributional semantics. In: *Proceedings of the 10th International Conference on Computational Semantics*. ; 2013:13-22.
88. Elhamifar E, Vidal R. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Trans Pattern Anal Mach Intell*. 2013;35(11):2765-2781.
89. Li D, Chaudhary H, Zhang Z. Modeling spatiotemporal pattern of depressive symptoms caused by COVID-19 using social media data mining. *Int J Environ Res Public Health*. 2020;17(14):4988.
90. Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ; 2014:1532-1543.
91. Naseem U, Razzak I, Musial K, Imran M. Transformer based deep intelligent contextual embedding for twitter sentiment analysis. *Future Generation Computer Systems*. 2020;113:58-69.
92. Burdisso SG, Errecalde M, Montes-y-Gómez M. A text classification framework for simple and effective early depression detection over social media streams. *Expert Syst Appl*. 2019;133:182-197.
93. Nguyen T, Phung D, Dao B, Venkatesh S, Berk M. Affective and content analysis of online depression communities. *IEEE Trans Affect Comput*. 2014;5(3):217-226.

94. Breiman L. Bagging predictors. *Mach Learn*. 1996;24(2):123-140.
95. Polikar R. Ensemble learning. In: *Ensemble Machine Learning*. Springer; 2012:1-34.
96. Skurichina M, Duin RPW. Stabilizing classifiers for very small sample sizes. In: *Proceedings of 13th International Conference on Pattern Recognition*. Vol 2. ; 1996:891-896.
97. Fatima I, Mukhtar H, Ahmad HF, Rajpoot K. Analysis of user-generated content from online social communities to characterise and predict depression degree. *J Inf Sci*. 2018;44(5):683-695.
98. Tabar M, Park H, Winkler S, Lee D, Barman-Adhikari A, Yadav A. Identifying homeless youth at-risk of substance use disorder: Data-driven insights for policymakers. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ; 2020:3092-3100.
99. Abbasi A, Li J, Clifford G, Taylor H. Make “fairness by design” part of machine learning. *Harv Bus Rev*. Published online 2018:2-7. https://hbr.org/2018/08/make-fairness-by-design-part-of-machine-learning
100. Lin H, Tov W, Qiu L. Emotional disclosure on social networking sites: The role of network structure and psychological needs. *Comput Human Behav*. 2014;41:342-350. doi:10.1016/j.chb.2014.09.045
101. Liu S, Zhu M, Yu DJ, Rasin A, Young SD. Using real-time social media technologies to monitor levels of perceived stress and emotional state in college students: A web-based questionnaire study. *JMIR Ment Health*. 2017;4(1):e2. doi:10.2196/mental.5626
102. De Choudhury M, Counts S, Horvitz E. Social media as a measurement tool of depression in populations. In: *Proceedings of the 5th Annual ACM Web Science Conference, WebSci’13*. ; 2013:47-56. doi:10.1145/2464464.2464480
103. Moreno MA, Christakis DA, Egan KG, et al. A pilot evaluation of associations between displayed depression references on facebook and self-reported depression using a clinical scale. *Journal of Behavioral Health Services and Research*. 2012;39(3):295-304. doi:10.1007/s11414-011-9258-7
104. Schwartz HA, Eichstaedt J, Kern ML, et al. Towards assessing changes in degree of depression through Facebook. In: *ACL Workshop on Computational Linguistics and Clinical Psychology*. ; 2014:118-125. doi:10.3115/v1/w14-3214
105. HHS Announces Nearly $35 Million To Strengthen Mental Health Support for Children and Young Adults. https://www.samhsa.gov/newsroom/press-announcements/20220309/hhs-announces-35-million-strengthen-mental-health
106. Liang H, Xue Y, Berger BA. Web-based intervention support system for health promotion. *Decis Support Syst*. 2006;42(1):435-449.
107. 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):708-713.
108. Solberg LB. Data mining on Facebook: A free space for researchers or an IRB nightmare? *U Ill JL Tech & Pol’y*. Published online 2010:311.

109. Kosinski M, Matz SC, Gosling SD, Popov V, Stillwell D. Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American psychologist*. 2015;70(6):543.
110. Dwyer C, Hiltz S, Passerini K. Trust and privacy concern within social networking sites: A comparison of Facebook and MySpace. *AMCIS 2007 proceedings*. Published online 2007:339.
111. Eubanks V. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press; 2018.
112. Marwick AE, Boyd D. Networked privacy: How teenagers negotiate context in social media. *New Media Soc*. 2014;16(7):1051-1067.
113. Boyd D, Crawford K. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Inf Commun Soc*. 2012;15(5):662-679.
114. Van Dijck J. Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveill Soc*. 2014;12(2):197-208.
115. Boyd DM, Ellison NB. Social network sites: Definition, history, and scholarship. *Journal of computer-mediated Communication*. 2007;13(1):210-230.
116. Rice E, Barman-Adhikari A. Internet and social media use as a resource among homeless youth. *Journal of Computer-Mediated Communication*. 2014;19(2):232-247. doi:10.1111/jcc4.12038
117. Bhandari A, Sun B. An online home for the homeless: A content analysis of the subreddit r/homeless. *New Media Soc*. 2023;25(9):2419-2436.
118. Park IY, Barman-Adhikari A, Shelton J, et al. Information and communication technologies use among youth experiencing homelessness: associations with online health information seeking behavior. *Inf Commun Soc*. Published online 2024:1-18.
119. Forchuk C, O'Regan T, Jeng M, Wright A. Retaining a sample of homeless youth. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*. 2018;27(3):167.

## Appendix A

Table A1. Summary of survey question categories with sample questions

| Categories | Sample Question | Number of Questions |
|---|---|---|
| Socio-demographic | What is your current gender identity? | 15 |
| HIV Testing and Treatment History | When was the last time you were tested for HIV/AIDS? | 25 |
| Healthcare Access and Utilization | Where do/would you prefer to receive health care services, in order of preference? | 23 |
| Homeless History | How many different times have you been without a stable place to stay? | 8 |
| Sexual-risk behaviors | How old were you when you had sexual intercourse for the first time? | 19 |
| Recent Substance Use | During the past 30 days, on how many days did you have at least one drink of alcohol | 25 |
| Technology Access and Use | Why do you typically use social networks?<br>In the past three months, when you are logged on to social networks, how much time do you spend communicating with (or trying to find) prospective sexual partners? | 52 |
| Mental Health Characteristics | Over the last 2 weeks, how often have you been bothered by the following problems? (This is the 9-item questionnaire that is used to assess the level of depression) | 28 |
| Victimization Experiences | In the past 3 months, has anyone hit or attacked you without using an object or weapon? | 29 |

Table A2. Substance questions and answer codes

| Substance | Survey Question | Answers |
|---|---|---|
| Marijuana | During the past 30 days, how many times did you use marijuana? | 1: 0 times; 2: 1 or 2 times; 3:3 to 9 times; 4: 10 to 19 times; 5: 20 to 39 times; 6: 40 or more times |
| Cocaine | During the past 30 days, how many times did you use any form of cocaine (including powder, coke, blow, or snow) but NOT crack? | Same as above |
| Crack | During the past 30 days, how many times did you use crack, including freebase or rock? | Same as above |
| Heroin | During the past 30 days, how many times have you used heroin (also called smack, junk, or China White)? | Same as above |
| Methamphetamines | During the past 30 days, how many times have you used methamphetamines (also called meth, speed, crystal, crank, or ice)? | Same as above |
| Ecstasy | During the past 30 days, how many times have you used ecstasy (also called MDMA or X)? | Save as above |

Table A3. Summary of substance use among participants

| Drug | Alcohol | Marijuana | Cocaine | Crack | Heroin | Meth | Ecstasy | Needle injections | Prescription drugs |
|---|---|---|---|---|---|---|---|---|---|
| **# of users** | 44 | 94 | 21 | 6 | 7 | 30 | 11 | 9 | 16 |
| **% of users** | 33.9 | 72.3 | 16.2 | 4.6 | 5.4 | 23.1 | 8.5 | 6.9 | 12.3 |

n=130, one missing value for methamphetamine

Table A4: Performance of VADER classification of 300 messages

| Metric | All Messages (300) | Positive Messages (129) | Negative Messages (86) | Neutral Messages (85) |
|---|---|---|---|---|
| **Sensitivity (Recall)** | 0.70 | 0.74 | 0.65 | 0.67 |
| **Precision** | 0.70 | 0.73 | 0.71 | 0.64 |
| **Specificity** | 0.85 | 0.79 | 0.89 | 0.85 |

Note: The numbers in parentheses represent the number of messages in each category. We used the micro-average approach to compute the overall recall, precision, and specificity.

Table A5. Summary of substance use among participants with active FB posts (our final data)

| Drug | Alcohol | Marijuana | Cocaine | Crack | Heroin | Meth | Ecstasy | Needle injections | Prescription drugs |
|---|---|---|---|---|---|---|---|---|---|
| **# of users** | 30 | 58 | 9 | 1 | 3 | 16 | 6 | 6 | 9 |
| **% of users** | 35.7 | 69.1 | 10.7 | 1.2 | 3.6 | 19.3 | 7.1 | 7.1 | 10.7 |

n=84, one missing value for methamphetamine

Table A6. Most contributing words for the top five latent topics

| No. | Top 10 Most Contributing Words | Latent Topic Theme | Distribution within User Group | Distribution within Non-user Group |
|---|---|---|---|---|
| #1 | really, find, tell, f**king, die, try, right, girl, happy, talk | Relationship | 38.9% | 37.5% |
| #2 | today, work, job, man, tomorrow, do, let, try, guy, place | Work | 22.9% | 21.5% |
| #3 | ass, b**h, tell, man, [N-word]s, let, even, talk, big, damn | Swear | 21.2% | 17.9% |
| #4 | lmaoo, baby, beautiful, sister, cute, today, morning, happy, girl, miss | Female-related | 5.1% | 6.2% |
| #5 | cause, real, world, keep, live, work, music, ill, soul, hard | Lifestyle | 5.4% | 4.3% |

Note: The topics are ordered by their occurrences in the posts. Keywords can belong to more than one topic, with different weights (probabilities). Words under the topic themes 1 and 3 are edited for the publication of this paper so as not to offend its readers.

Table A7: Accuracy of models using different features

| | Feature Used | | | |
|---|---|---|---|---|
| | Accuracy | | AUC | |
| Model | Five Feature Sets with Word Embeddings | Survey Information | Five Feature Sets with Word Embeddings | Survey Information |
| NN+ Bagged Decision Tree | 0.81 | 0.56 | 0.72 | 0.44 |
| NN+ Bagged SVC | 0.69 | 0.56 | 0.50 | 0.50 |
| NN+ Bagged Logit | 0.76 | 0.69 | 0.66 | 0.52 |

Note: Because the survey information does not contain word embeddings, NN was not used to leverage word embeddings. Survey information contains the following: age, perceived health, whether the participant is working or not, whether the participant attends school or not, has the participant been in jail or not, education level, has the participant been attacked or not, race, gender, level of depression, and level of anxiety

Table A8. Summary of marijuana use by gender and age

| | Marijuana User | Non-user | Total # of Observations |
|---|---|---|---|
| **Group by Gender** | | | |
| **Male** | 0.73 | 0.27 | 49 |
| **Female** | 0.67 | 0.33 | 27 |
| **Other** | 0.5 | 0.5 | 8 |
| **Group by Age** | | | |
| **Below 21** | 0.68 | 0.32 | 47 |
| **21 and Above** | 0.70 | 0.30 | 37 |

n=84

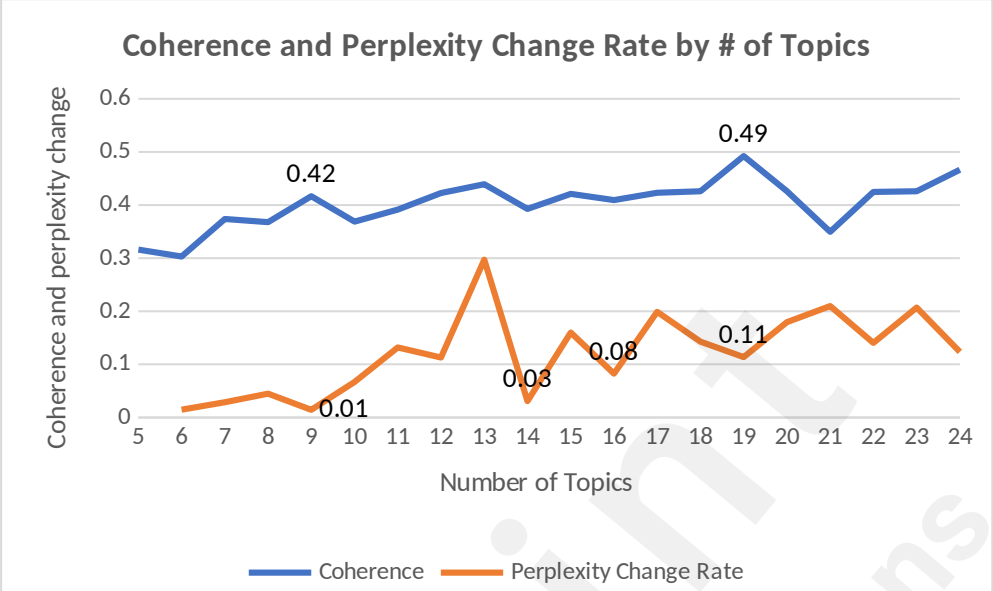



**Figure A1. Coherence and Perplexity Change Rate by Number of Topics**

## Appendix B

We describe the process through which VADER calculates and normalizes the sentiment score of a piece of text.

VADER provides the following four scores:

1. Negative Score: This score represents the proportion of text that conveys negative sentiment. It is a measure of the amount of negative sentiment words present in the text.

2. Positive Score: This score represents the proportion of text that conveys positive sentiment. It measures the amount of positive sentiment words in the text.

3. Neutral Score: This score represents the proportion of text that is neutral, meaning it does not convey strong sentiment in either direction (positive or negative). It measures the amount of text that is neither positive nor negative.

4. In this study, we used the compound score, which is the overall sentiment of the text. The compound score is a normalized, weighted composite score that ranges from -1 (most negative) to +1 (most positive). It is calculated by summing the valence scores of each lexicon (including words, phrases, punctuations, and emojis) in the text, taking account of grammatical and syntactical rules such as negation and degree intensifiers. Then VADER uses the following formula to normalize this overall sentiment score:

$$Compound\ Score = \frac{\sum of\ all\ valence\ scores}{\sqrt{\sum of\ valence\ scores^2 + \alpha}}$$

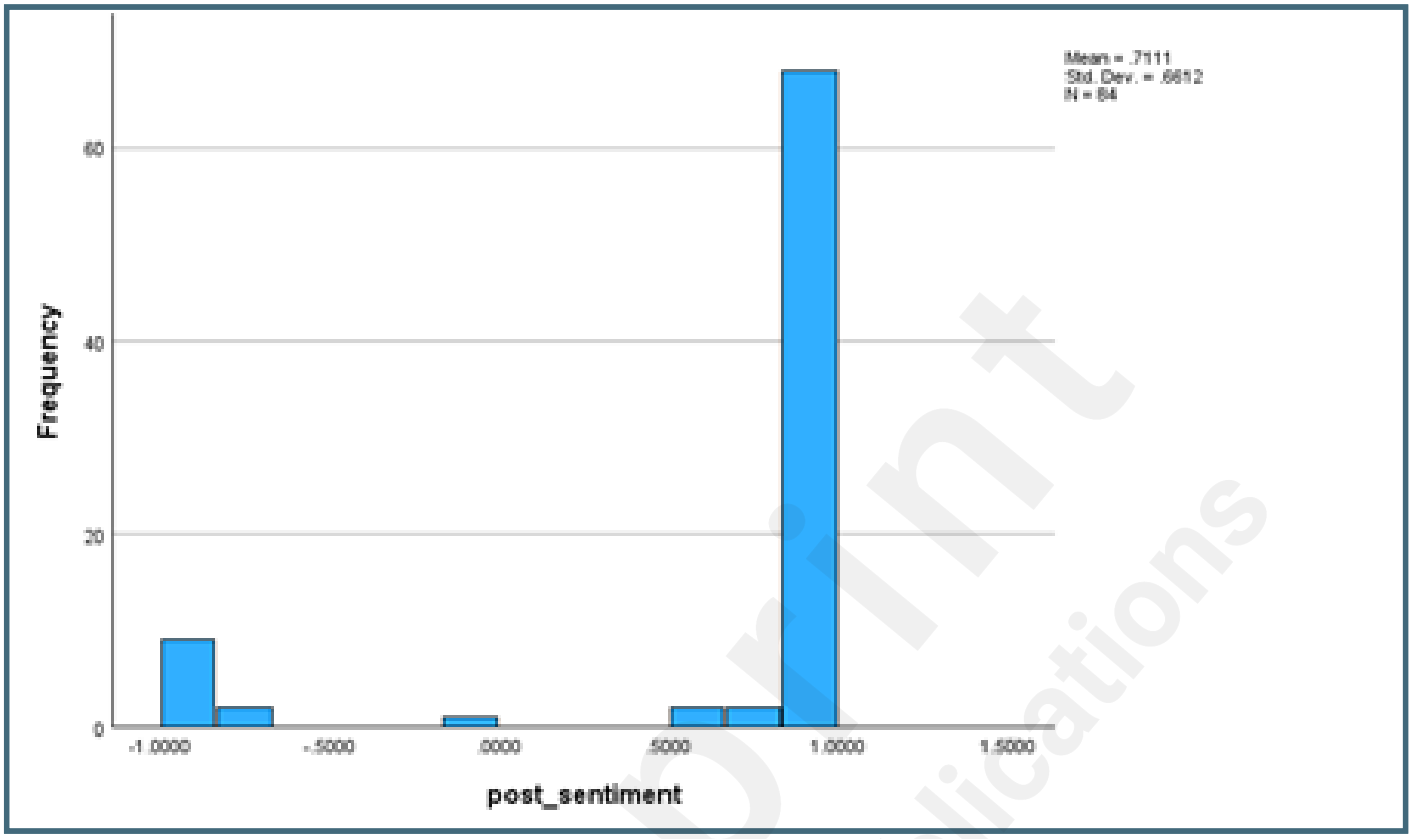
α is a normalization constant that ensures the scores fall within the range of -1 to 1. VADER uses a value of 15 for $\alpha$, which helps scale the scores appropriately. By following these steps, VADER can provide a normalized sentiment score that ranges from -1 (most negative) to 1 (most positive). This compound score gives an overall sentiment of the text based on the individual word valences and their contextual adjustments. It is often used as a single metric to determine the general sentiment of the text.

# Supplementary Files

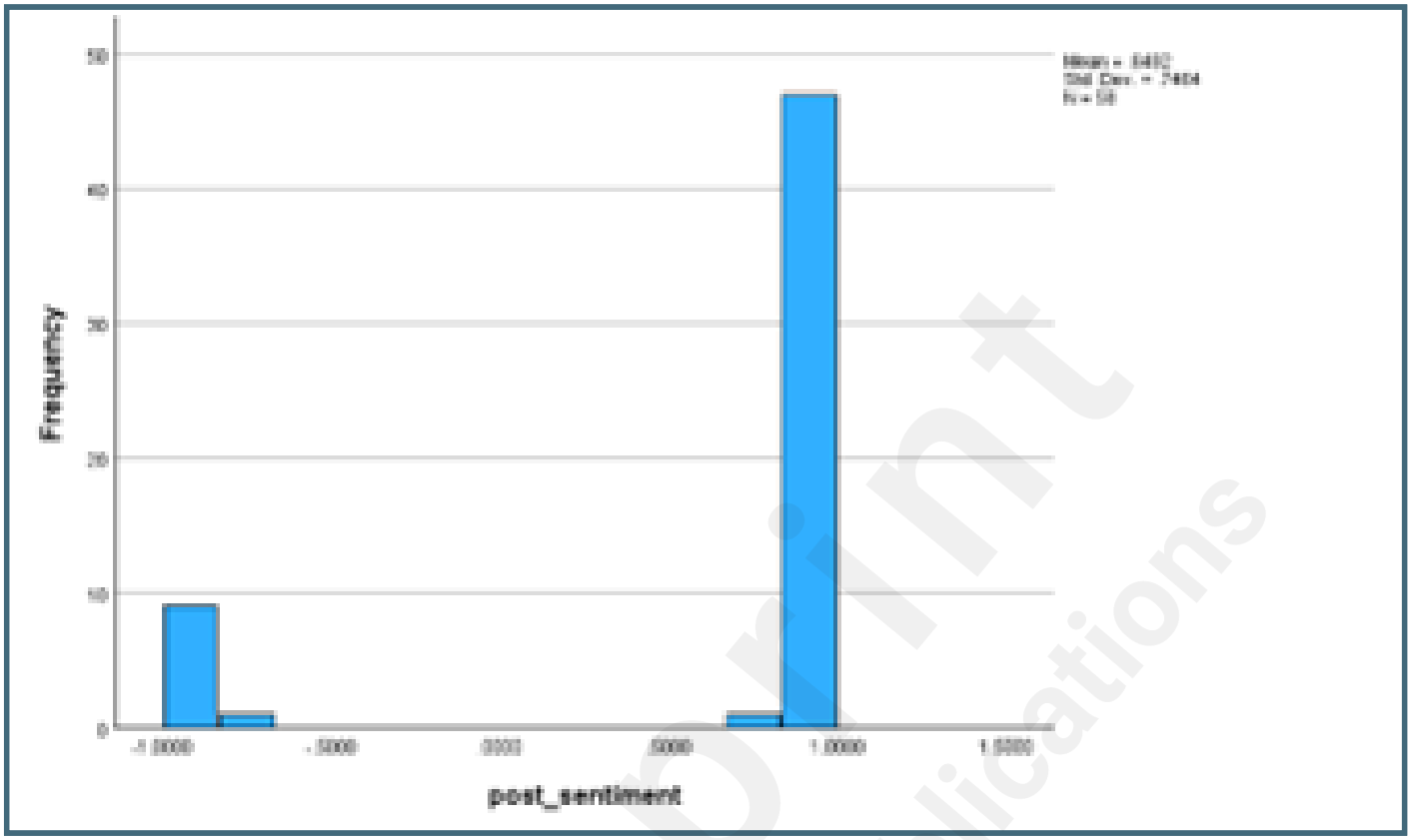
Untitled.
URL: http://asset.jmir.pub/assets/cb53607078a6b90d9a2bbac070edbf81.docx

# Figures

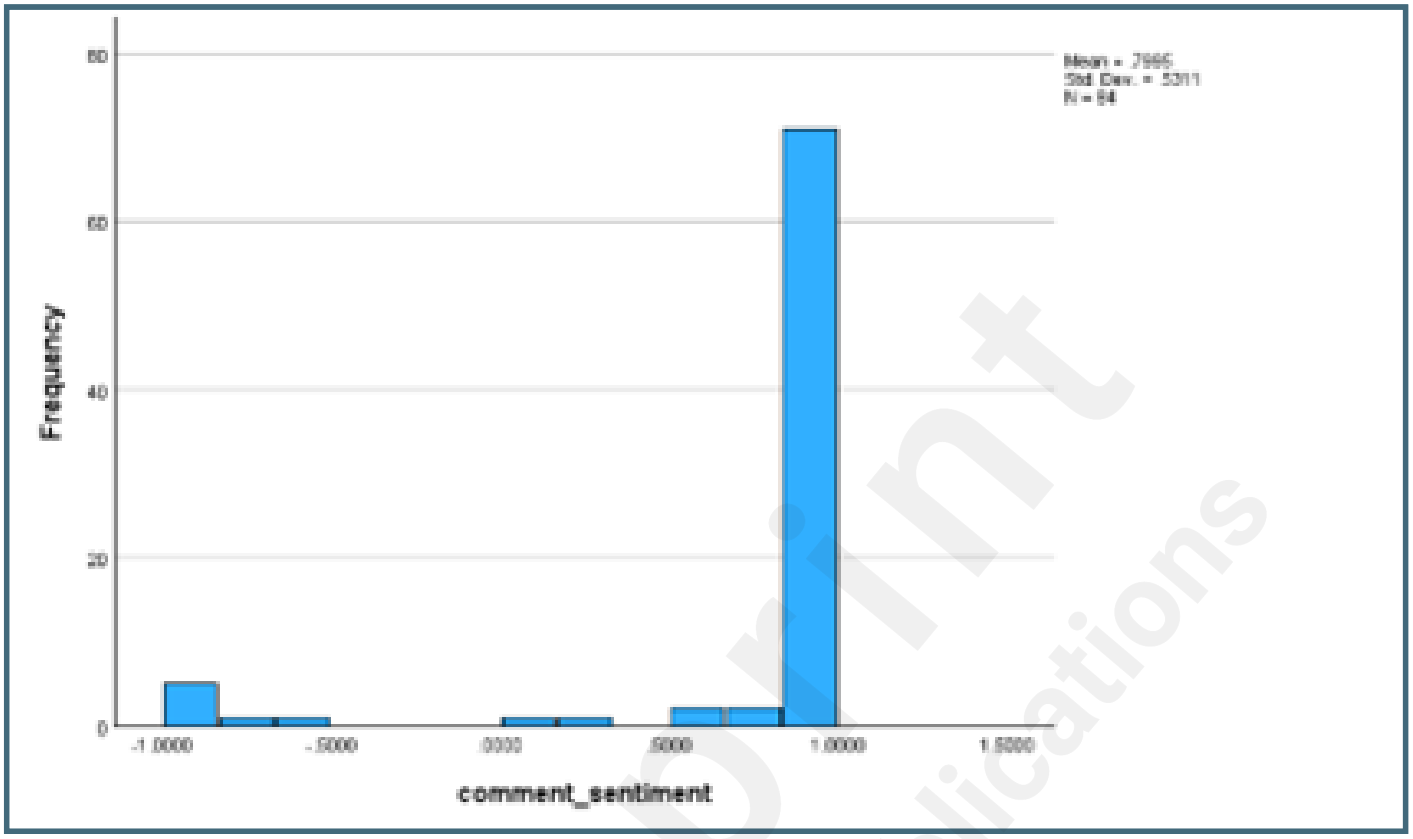
Untitled.



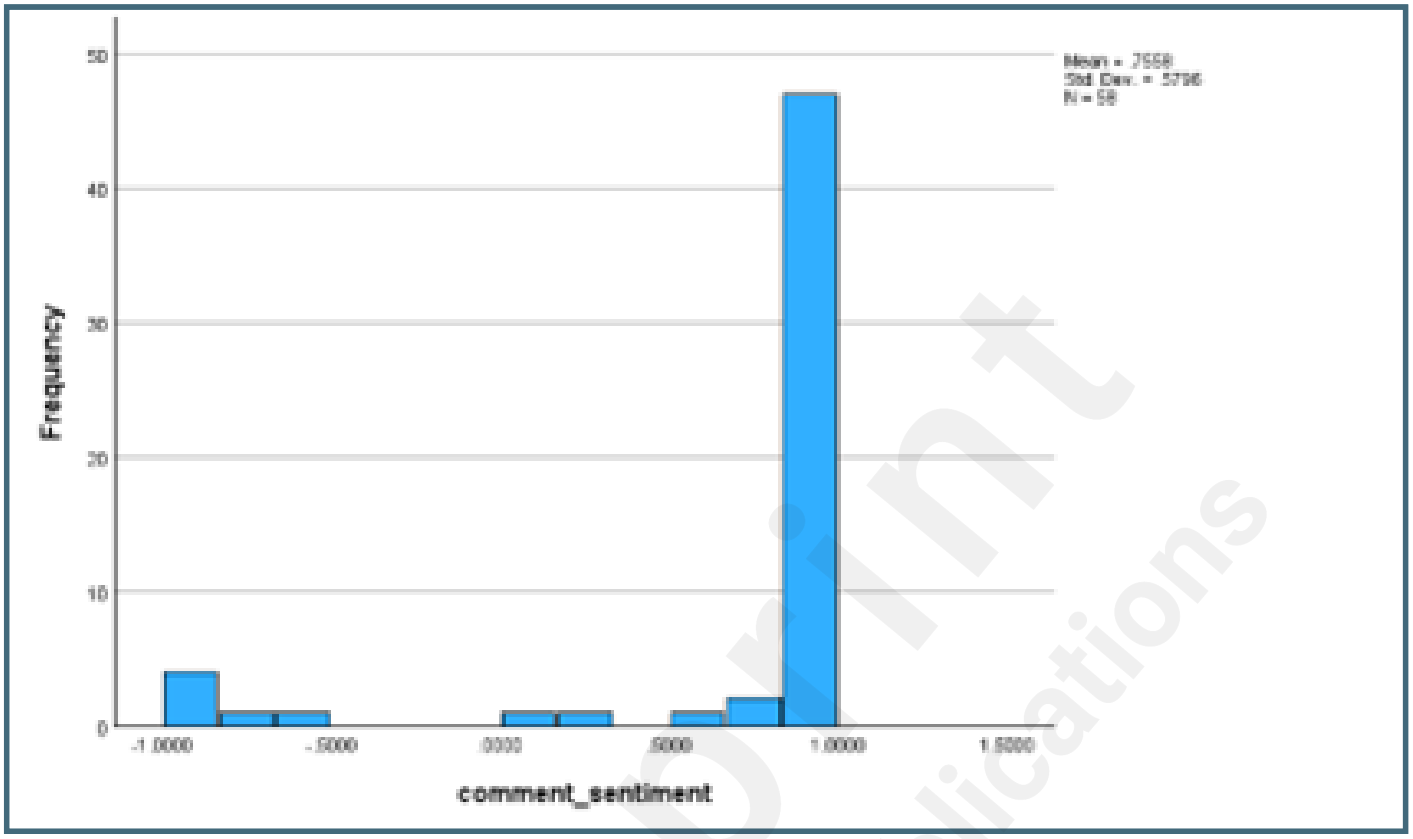
Untitled.



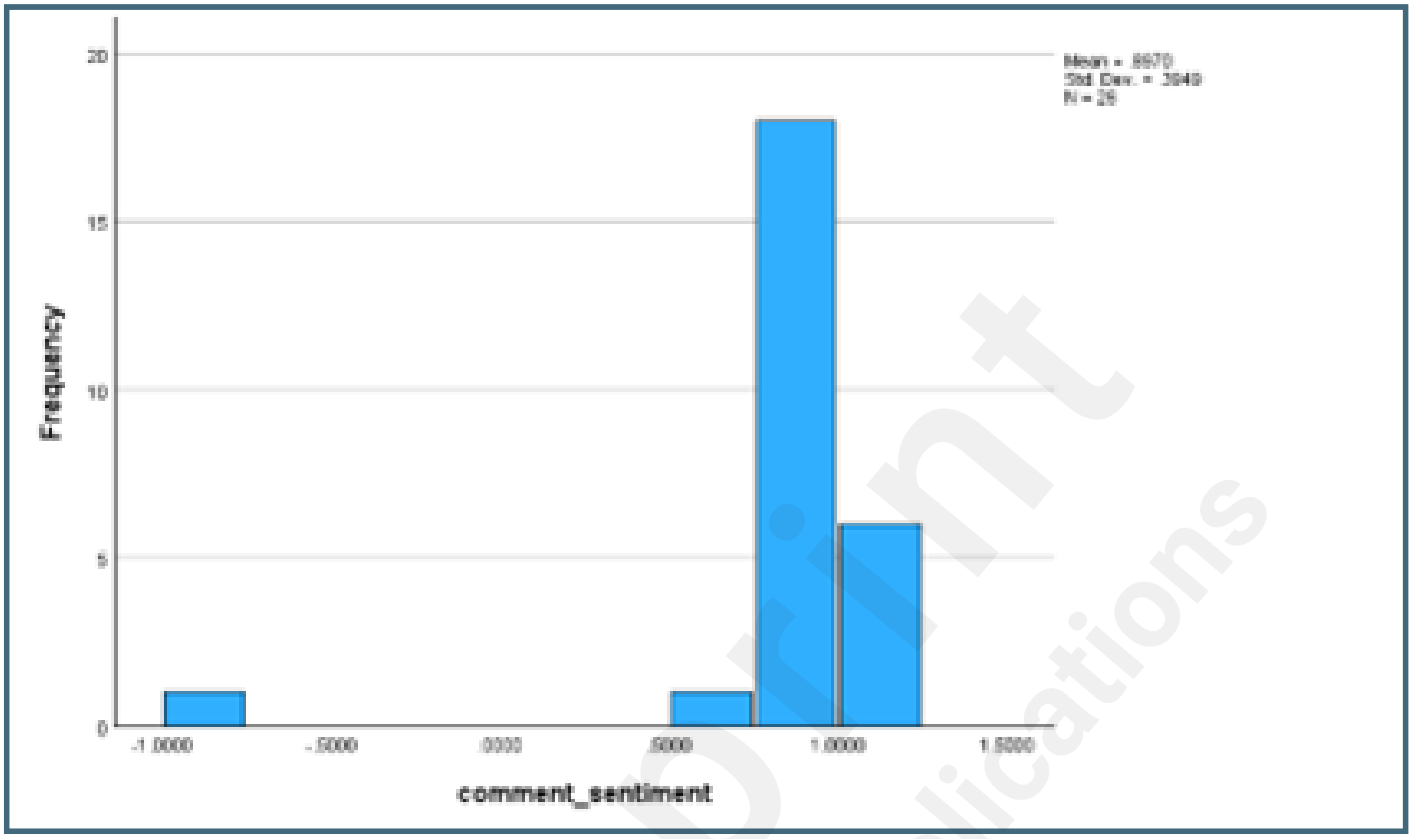
Untitled.



Untitled.



Untitled.



Sentiment distribution non-user.