

Decoding the Contexts: A Natural Language Processing Analysis of Peer and Family Dynamics in Teenage Substance Use Online Discussions

Jianfeng Zhu, Xinyu Zhang, Ruoming Jin, Deric R Kenne

Submitted to: Journal of Medical Internet Research
on: May 28, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript.....	4
--------------------------	---

Preprint
JMIR Publications

Decoding the Contexts: A Natural Language Processing Analysis of Peer and Family Dynamics in Teenage Substance Use Online Discussions

Jianfeng Zhu¹ PhD; Xinyu Zhang² PhD; Ruoming Jin³ Prof Dr; Deric R Kenne⁴ Prof Dr

¹Department of Communication, Michigan State University East Lansing US

²Department of Computer Science, Kent State University Kent US

³Center for Public Policy and Health, College of Public Health, Kent State University Kent US

Corresponding Author:

Jianfeng Zhu PhD

Abstract

Background: Substance use among adolescents represents a significant public health concern. Despite research on the influence of contextual factors on teenage substance use, the nuances of these contexts and associated emotions remain underexplored, particularly within online communities.

Objective: This study aims to decode the context factors, emotion tones, and topics discussed in substance use-related posts within the r/teenager subreddit. We introduce an innovative approach leveraging natural language processing to analyze discussions within a large online community of teenagers on Reddit.

Methods: We collected 31,724 substance use-related posts from the r/teenagers subreddit in English, spanning from 2018 to 2023. A mixed methods was applied on these posts, including trend analysis to explore the dynamic context within the data. The BERT family of models was used to extract sentiment and topic features.

Results: The analysis revealed that "drink" and "weed" were the most frequently mentioned substances among teenagers. Peaks in discussions occurred during summer months, particularly in July 2023. Substance use posts exhibited higher levels of negative emotions such as annoyance, sadness, and disappointment. Family dynamics significantly impacted the teenage substance use environment and behaviors. The term "Dad" was commonly associated with familial roles and the setting of examples within the household, particularly regarding habits and behaviors around substance consumption and addiction. The term "Mum" was often at the forefront of familial discussions about marijuana and smoking cessation.

Conclusions: The study's findings provide deeper insights into the multifaceted nature of substance use discussions among teenagers, informing potential intervention strategies and future research directions. The authors underscore the need for targeted interventions addressing peer influence, familial communication, and emotional well-being in mitigating adolescent substance use.

(JMIR Preprints 28/05/2024:60924)

DOI: <https://doi.org/10.2196/preprints.60924>

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

Decoding the Contexts: A Natural Language Processing Analysis of Peer and Family Dynamics in Teenage Substance Use Online Discussions

Abstract

Background: Substance use among adolescents represents a significant public health concern. Despite research on the influence of contextual factors on teenage substance use, the nuances of these contexts and associated emotions remain underexplored, particularly within online communities.

Objective: This study aims to decode the context factors, emotion tones, and topics discussed in substance use-related posts within the r/teenager subreddit. We introduce an innovative approach leveraging natural language processing to analyze discussions within a large online community of teenagers on Reddit.

Methods: We collected 31,724 substance use-related posts from the r/teenagers subreddit in English, spanning from 2018 to 2023. A mixed methods was applied on these posts, including trend analysis to explore the dynamic context within the data. The BERT family of models was used to extract sentiment and topic features.

Result: The analysis revealed that "drink" and "weed" were the most frequently mentioned substances among teenagers. Peaks in discussions occurred during summer months, particularly in July 2023. Substance use posts exhibited higher levels of negative emotions such as annoyance, sadness, and disappointment. Family dynamics significantly impacted the teenage substance use environment and behaviors. The term "Dad" was commonly associated with familial roles and the setting of examples within the household, particularly regarding habits and behaviors around substance consumption and addiction. The term "Mum" was often at the forefront of familial discussions about marijuana and smoking cessation.

Conclusion: The study's findings provide deeper insights into the multifaceted nature of substance use discussions among teenagers, informing potential intervention strategies and future research directions. The authors underscore the need for targeted interventions addressing peer influence, familial communication, and emotional well-being in mitigating adolescent substance use.

Keywords: Adolescents, Natural language processing, Contextual factors, Reddit, Substance use

Introduction

Substance use among adolescents and young adults represents a significant public health concern, particularly given the propensity for binge drinking and drug use within this demographic. Despite accounting for only 3.2% of all alcohol consumed in the United States, individuals aged 12 to 20 often engage in consumption patterns that result in higher intake per occasion compared to adults [1]. This behavior underscores a critical health burden, highlighting the urgent need for effective interventions and policy measures tailored to this age group [2].

Adolescent substance use is a complex issue, influenced by genetics, environment, peer relationships, and specific contexts, varying from experimental to pathological use. This diversity underlines the multifaceted nature of substance use among teenagers, with each context presenting unique risks [3]. The decision-making process regarding substance use involves both

physical (e.g., privacy, safety) and social aspects (e.g., peer influence, adult supervision), affecting adolescents' perceptions and behaviors differently in public versus private settings [4]. Despite the recognized complexity and the significant impact of context on substance use patterns, research specifically addressing the nuances of these contexts and the emotions associated with different settings is scarce. This gap highlights the need for focused studies on how various contexts influence adolescents' substance use decisions and the associated emotional experiences.

In response to this gap, our study introduces an innovative approach by leveraging natural language processing (NLP) to analyze discussions within a large online community of teenagers. This method allows for an unprecedented examination of the various contexts surrounding substance use, including where and how these behaviors are most likely to unfold. By focusing on the situational and environmental factors discussed by adolescents themselves, we gain insights into the specific circumstances that may encourage or discourage substance use. This perspective is crucial for developing more nuanced and effective intervention strategies that can be tailored to address the real-world complexities faced by young individuals.

Literature review

Teenagers' Substance Use

The passage emphasizes the initial phase of substance use among teenagers as a period of experimentation, which holds the potential for developing into a dependency on substances like tobacco, alcohol, or marijuana. This stage is crucial because adolescents have not yet fully committed to substance use, indicating a window of opportunity for effective interventions to prevent the escalation into regular use [5]. To design these interventions effectively, it's vital to comprehend teenagers' perceptions of substance use, as these views form the basis for their future behavior and attitudes towards these substances.

Limited research focused on teenagers' substance use. It's found that Teenager may turn to substance for various reasons, including curiosity and peer pressure, social conformity, relaxation, boredom alleviation, pain management, coping with trauma, anxiety relief, and weight control [6]. Substance use often becomes reinforcing when it appears to alleviate specific problems or needs.

Diagnosing substance use and misuse in teenagers can be challenging due to lack of information. It is critical for clinicians to get a full picture of what symptoms occurred first and when substance misuse began. Emotional and Behavioral changes, such as moodiness, insomnia, avoiding friends and lack of energy, which those can result from both mental health disorders and substance use [7].

Contextual factors

Contextual factors play a critical role in shaping teenagers' substance use, affecting not only how often they use substances but also the types of substances they choose. These factors encompass a wide range of environmental, social, and individual elements that interact in complex ways to influence behavior. Understanding these contexts is vital for two main reasons.

Firstly, by understanding the environmental, social, and individual contexts in which substance use occurs, we can identify specific risk and protective factors at play [8]. For example, interventions that aim to strengthen family bonds and improve parental supervision may be more effective in environments where these elements are identified as lacking. Furthermore, focusing on contextual factors facilitates a deeper understanding of the mechanisms through which substance use behaviors are learned, reinforced, or mitigated, guiding the development of theoretical models that can predict behavior change [9]. Research highlights the importance of family dynamics, peer influence, school and neighborhood environment as key areas of focus, underscoring the complex interplay of influences that shape adolescent substance use.

Family dynamics significantly influence adolescent substance use. Authoritarian parenting, marked by high demands and low responsiveness, increases the likelihood of teenagers smoking. Conversely, permissive parenting, characterized by warmth but minimal control, is linked to lower rates of adolescent drinking. Neglectful parenting, with low support and control, correlates with higher rates of substance experimentation [10]. Family conflicts further heighten the risk of substance use [10]. Additionally, parental attitudes towards alcohol and their own usage patterns shape adolescent behaviors; negative views on alcohol by parents are associated with lower adolescent drinking levels [11]. Moreover, alcohol-related issues within the family, particularly among parents, significantly elevate adolescents' risk of substance abuse, emphasizing the critical role of parental behavior and family environment in shaping adolescent substance use patterns [10].

Peer influence significantly shapes adolescent substance use. Teens are more inclined to use substances if their friends do, driven by the perception of peers' behaviors and the desire to fit in [10]. Peer pressure, along with the fear of social exclusion or rejection [12], can exacerbate emotional and behavioral health issues, prompting adolescents to use substances as a coping mechanism or to seek acceptance in alternative peer groups [13]. Furthermore, experiences of social exclusion or peer rejection can lead to emotional and behavioral health issues, including increased likelihood of substance use. Adolescents may turn to substances as a coping mechanism or attempt to gain acceptance in other peer groups [13]. Research also highlights the role of the immediate environment in facilitating peer influence, with substance use more prevalent in smaller group settings. A direct correlation exists between the extent of friends' substance use and an individual's own usage, with increases in friends' consumption at social gatherings like parties leading to corresponding increases in personal substance use [9]. Similarly, Egan et al. [14] reveal that a significant portion of parties attended by 15-20-year-olds feature not just alcohol but also marijuana and/or illicit drug use, underlining the social contexts that fuel substance use among adolescents. Targeted interventions that address these peer and environmental influences are crucial for mitigating substance use in this demographic.

Additionally, research indicates that substance use among students significantly varies across schools, suggesting the school environment plays a crucial role in shaping students' behaviors towards substances like alcohol and illicit drugs [15]. Schools with cultures that tacitly or openly condone substance use often

report higher instances of such behaviors. Conversely, environments promoting academic excellence can paradoxically increase certain risk behaviors [16]; for instance, high academic achievement has been linked to smoking and binge drinking [17]. Additionally, the quality of student-teacher interactions and overall satisfaction with school life play roles in substance use patterns, with effective communication and contentment being predictors of marijuana use [17]. However, students facing academic challenges or feeling detached from their educational institution are more susceptible to substance use, underlining the need for supportive academic environments.

Also, the relationship between neighborhood context and teenager's substance use has been explored in various studies, highlighting the significant influence of neighborhood characteristics on adolescent behavior. A study found that perceptions of neighborhood disorganization (e.g., violence/safety and drug activity) in grade 7 were associated with increased tobacco, alcohol, and marijuana use in grade 9 among urban African American adolescents. For females, this association was mediated by attitudes about drug use and perceptions of drug harmfulness, emphasizing the importance of neighborhood contextual variables for substance use among African American adolescents [8]. Another study explored the additive and moderating effects of neighborhood context on substance use among youth. It was found that adolescents living in poor neighborhoods have lower rates of alcohol and marijuana use. Additionally, neighborhood disadvantage moderated the effect of deviant peers on adolescent substance use, suggesting that living in areas with higher neighborhood disadvantage reduces the effect of deviant peers on substance use [18]. Similarly, it was suggested that the relationship between victimization and marijuana use was stronger for those in neighborhoods with greater disadvantage [19].

Collectively, these studies have respectively underscored the critical role of family dynamics, peer influence, school, and neighborhood environment in shaping adolescent substance use. Yet, the detailed interplay and relative importance of these contextual factors remain underexplored, particularly in how specific topics within these areas influence behavior. Our research aims to bridge this gap by adopting a novel approach that focuses on online communities. By harnessing naturalistic data from online communities—spaces where adolescents candidly share thoughts and experiences—we aim to dissect the intricate interplay between digital and traditional socialization. This novel approach allows for a detailed examination of how specific topics within these contextual factors influence adolescent behavior.

Research Hypothesis Questions and Objectives

To expand upon the limited research examining the relationship between teen substance use and contextual factors, the goal of current study was to investigate the substance use and contextual environment in online communities through r/Teenagers subreddit from 2018 to 2023. Additionally, this study attempts to characterize the substance use posts through the prevalence of emotional tone expression, contextual factors, and emerging topics, while exploring correlations. We sought to address the following research questions:

RQ1: What are the most emotional features within teen substance use posts?

RQ2: What are the important contextual factors popular in the discussion?

RQ3: What are the most prominent topics among these posts?

RQ4: What correlations exist among emotional, contextual factor, and topics?

Methods

Data Collection

The data of interest were collected from Reddit, an online forum known as a home to thousands of communities and authentic human connection. Posts made on Reddit are organized into communities of interest and content known as subreddits [20]. This study focuses on posts in the r/teenagers subreddit, created in 2010, which is the biggest community forum run by teenagers for teenagers. As of March 2024, it had over 3.0 million members, targeting users aged between 13 and 19. A historic archive of posts made in the teenagers subreddit was collected using the publicly available site Internet Archive[21]. The website offers free and unlimited access to all posts, comments of users on Reddit. The full set of teenager's posts (N= 21,459,97) spanning from 2018 to 2023 was collected. Reddit posts that have been deleted by the user or removed through moderation were filtered from the dataset, as well as those posts with no value in the selftext field. Since we are interested in the original posts by users, so we exclude the comments in this study.

Reddit Data Preprocessing and Extraction

In our study, our aim was to identify posts related to substance use among teenagers by employing a comprehensive list of keywords associated with common substances such as alcohol, tobacco, marijuana, hallucinogens, sedatives, stimulants, and opioids. Our initial analysis revealed that these keywords also appeared in titles but not in the main posts. This realization prompted us to refine our methodology by combining 'title' and 'selftext' fields to capture all pertinent content, thereby enabling us to identify a greater number of posts related to substance use. To accurately extract relevant information, we used a specific regex pattern that ensured precise matching at word boundaries, leading to the identification of 31,724 substance-associated posts. To validate the effectiveness of our refined approach, we conducted a manual review of 300 randomly selected posts, which were evaluated by four workers. This validation process confirmed a 90% accuracy rate in identifying relevant posts. Table 1 represents samples of substance use-related posts with the matched terms.

Table 1 Sample of substance use-related posts with the matched terms

Filtered Posts	Matched Term
i'm allowed to take 1-2mg of xanax when i get a severe panic attack. but even the xanax doesn't work that well. i'm so unhappy and i can't even visit my family anymore	Xanax
Suicidal thoughts i always have em every	Weed, Alcohol

second im awake. no amount of weed or alcohol or nicotine or other drugs help.	
Got my first real shooting threat yesterday my school has reached peak trash. there was a fight between guy 1 and guy 2, guy 1 was sold a broken \$20 vape from guy 2, as guy 1 was being dragged off him he shouted "i got guns, too! i'll fucking prove it tomorrow!"\n\nimagine your friends being killed because someone was angry they got sold a broken \$20 vape.	Vape
Bro some people got in a fight at school so I went home early i also had like so many more cigarettes than i usually do in an entire month, holy fucking shit	cigarettes

In order to gain a better understanding of adolescent post-substance use behavior, we also collected an equal amount of data on non-substance use posts as baseline. This step was essential for broadening the scope of our analysis, allowing us to delve more deeper into the emotions, topics, and contexts prevalent in online discourse. By incorporating these non-substances use posts, we laid the groundwork for conducting a comparative examination aimed to identifying import features and themes that differentiate discussions related to substance use from other topics. This comparative approach enriches our understanding of the online landscape, providing valuable insights into the broader context surrounding discussions on substance use among teenagers.

Contextual Analysis

The context of a group is believed to influence the actions and attitudes of its members. Contextual analysis is the study of the role of the group context on actions and attitudes of individuals [22]. In our study, we applied contextual analysis to investigate the various contexts surrounding substance use-related context posted by teenagers. By observing the prevalence and content of different contexts mentioned in the posts over time, we aim to gain insights into the factors influencing substance use among adolescents.

To conduct the analysis, we first identified terms related to the family, peer interactions, school, and neighborhood contextual from previews research papers [8-19]. Additionally, these terms were refined by examine their frequency within dataset, with terms having a frequency greater than 100 being used in the contextual word analysis. For example, "parents", "family", and "mother" are high frequency word for family context, "friends" and "group" for peer contextual, "school" and "class" for school context. All terms related to neighborhood contextual showed lower frequency compared to the other three contexts.

Using these predefined context-related terms, we analyze the frequency of occurrence of each contextual factor in substance use-related posts. Furthermore, we measured the importance contextual factor influencing substance use posts using regress models. Additionally, we examined emotions and themes of substance use posts in relations to family, peer, school, and

neighborhood contextual factors.

Emotional Analysis

Emotions play a import role in conveying information within a particular context. In this study, we utilized the Roberta-base-go_emotions for emotional analysis. The model was trained on a dataset comprising approximately 58k manually annotated English Reddit comments, labeled for 27 emotions categories plus Neutral[23]. The GoEmotions dataset serves as a valuable resource for training and evaluating machine learning models designed to understand and interpret emotions expressed in text data.

By incorporating the GoEmotions model to extract emotion features from substance use posts, we aim to provide a comprehensive understanding of the nuanced emotional expression found in teenagers substance use-related content across different contextual or themes. This approach allows for a deeper exploration of the emotional aspects of substance use discussions, shedding light on complex interplay between emotions and contextual factors.

BERTopic Analysis

Topic modeling is a widely used text analysis tool for extracting latent variables from shared data points. In the context of text analysis, this typically involves identifying the events or concepts discussed in a document [24]. In our study, we utilized BERTopic, a state-of-the-art topic modeling technique that leverages pre-trained language embedding and a class-based TF-IDF approach to generate dense clusters. This method enables the creation of easily interpretable topics while preserving important words in the topics descriptions. BERTopic produces more accurate and contextually rich representations compared to traditional topic modeling techniques like Latent Dirichlet Allocation (LDA) [25].

Moreover, BERTopic offers advanced features such as Zero-Shot Topic Modeling, which allows users to discover predefined topics in large amounts of documents using semantic search. Additionally, BERTopic supports model merging, enabling the combination of multiple pre-trained BERTopic models to create a comprehensive topic model [26]. By employing these advanced techniques, our study facilitates a deeper exploration of specific topics and contextual terms related to substance use, thereby providing a comprehensive understanding of adolescents' concerns and perceptions.

Feature importance and correlation Analysis

Feature importance refer to techniques that calculate a score for all the input features in a given model, representing the “importance” of each feature. A high score indicates that the specific feature will have a larger effect on the model used to predict a certain variable [27]. We applied important features using a Random Forest classifier to predict substance use posts. This approach enabled us to identify the most impactful elements, including emotional features and contextual factors.

The correlation coefficient is a statistical measure of the strength of linear relationship between two variables. Its values can range from -1 to 1. A correlation coefficient of -1 describes a negative correlation, where values in one series rise as those in the other decline, and vice versa. To assess the degree of association between emotional features, contextual factor, and topics distribution within the substance use dataset, we utilized the panda DataFrame correlation function to compute pairwise correlation between generated features and factors. The function used Pearson correlation by default [28].

This comprehensive methodological approach not only underscored the intricate

interplay between various social environments and emotional responses in substance use discourse but also brought to light the prevalent underlying themes within the discussions.

Ethical Considerations

In our study, we strictly followed ethical guidelines using anonymized public data to protect privacy. We preprocessed the data by removing stopwords and applying lemmatization, ensuring no personal identifiers were included.

Results

Substance use posts distribution

Figure 1 represents the distribution of substance use-related posts in r/Teenagers subreddit community. The peak time occurs July 2023. As for the distribution of substance use terms, 'drink' accounts for 10,179 out of 32,803 posts (31%), 'weed' for 9,181 out of 32,803 posts (28%), 'alcohol' for 5398 out of 32,803 posts (16%), 'vape' for 2991 out of 32,803 (9%), and 'cigarettes' for 1,395 out of 32,803 posts (4%), making it the most frequently mentioned substance.

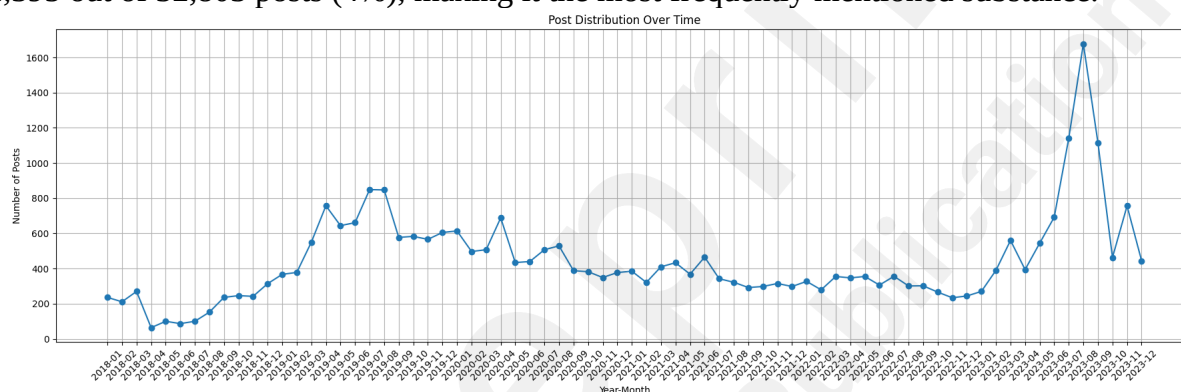


Figure 1

Substance use posts distribution from 2018 to 2023

Emotional analysis

Overall, we computed 27 emotional features from both substance use-related posts and the baseline dataset. All features were aggregated to the mean value. Figure 2 depicts the emotional mean values for substance use posts and normal posts. In the radar chart, each axis presents a different emotion, and the length of the line along each axis indicates the mean value of that emotion, ranging from 0.1 to 0.5. The red shape represents substance use posts, while the blue shape represents no substance use posts.

For substance use posts, the top five mean emotions are annoyance (0.46), approval (0.41), sadness (0.38), disappointment (0.31), and confusion (0.30). In contrast, the top five mean value, for non-substance use posts are annoyance (0.19), curiosity (0.18), approval (0.16), confusion (0.15), and sadness (0.14).

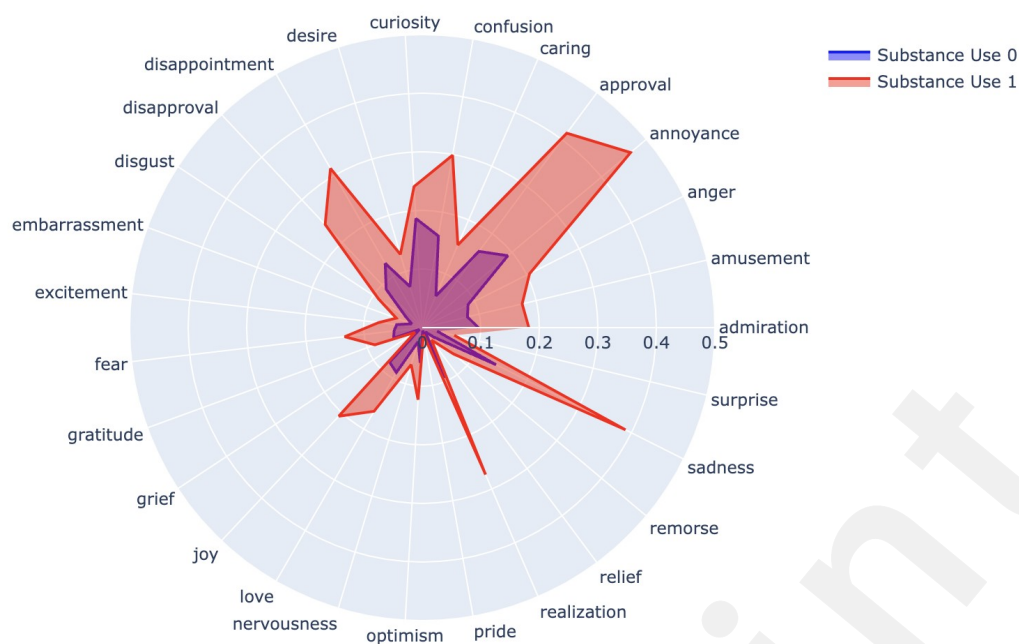


Figure 2 Radar Chart of mean emotion values for substance use categories posts

Contextual analysis

The analysis indicates that the peer context is prominent among the four contexts. "Friends" is the most prevalent term with a frequency 13,904 within the peer context, which totals 21, 252 mentions, suggesting that discussions about substance use frequently involve peer influence and interactions. The school context is the second most prominent, with the term "school" surpassing 12,000 mentions, this may feature the importance of the educational environment in the conversations. The family context follows, with "parents" being the most frequently mentioned term, indicating that discussions about substance use relate to parental figures. Other common terms include "dad" and "home", which suggests that family dynamics and home environments play roles in discussions about substance use among teenagers. The neighborhood context is less notable, with the most frequency term "street" having fewer than 600 mentions.

Figures 3 depict the distribution of each contextual factor from 2018 to 2023. The peer and family context lines show fluctuate over time but maintain relatively higher values compared to the school and neighborhood context. The peer context maintains the highest values in 2023, while the family context peaks during the pandemic, coinciding with school lockdowns due to the COVID-19. This suggests that increased time at home may have influenced the prevalence of family context discussions. The school context shows a consistent pattern with fluctuations, including a sharp decrease around March 2020, stabilizing at a lower level, and then returning to pre-pandemic levels around September 2021.

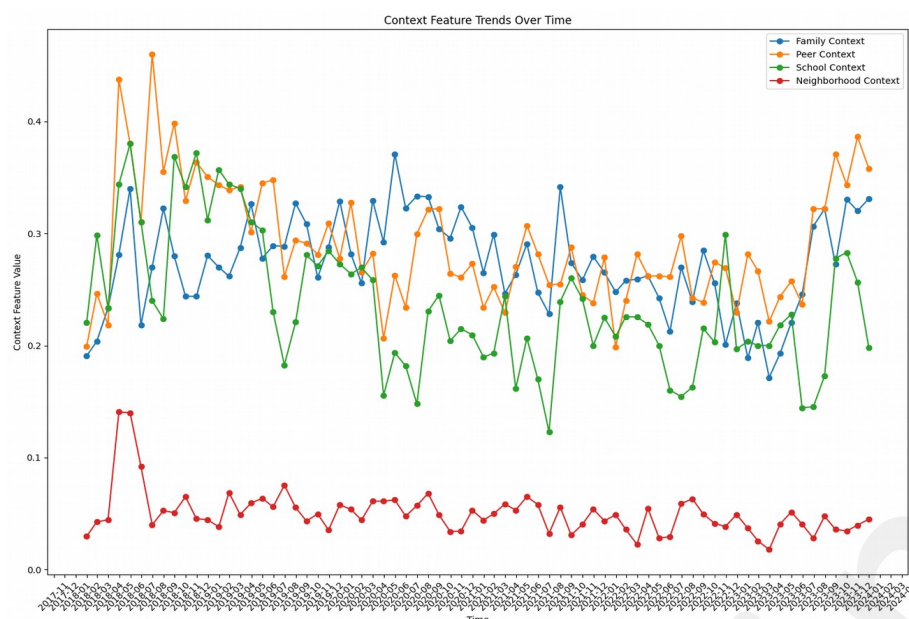


Figure 3 Contextual factors trend over time

BERTopic analysis

Table 2 displays the 16 topics identified through BERTopic, with ‘count’ indicating the number of documents associated with each topic in the dataset. KeyBERT, a keyword extraction method that leverages BERT embeddings, generates keyphrases closely resembling the content of a document’s [29]. ‘Custom Label’ refer to labels manually assigned by our team after reviewing the keywords and related documents for each topic.

Among the total (n=32,484) posts, the largest topic, comprising 61.7% (20,055), explores the interplay of identity and relationships. The discussion of diverse sensory perceptions of alcohol consumption total 7,393 posts, account for 23% of dataset. Topics such as discussing marijuana within family dynamics and parental influence, cocaine use, party culture, and the risk and consequences of vaping and nicotine amongst youth, collectively represent 2% of the dataset. The remaining topics cover a broad range of experiences, including substance use abroad, music, mental health, struggles, and social relationships related to substance use.

Table 2 BERTopic representations

Topic	Count	KeyBERT	Custom Topic Label
0	20066	['dating', 'crush', 'drunk', 'girl', 'best friend', 'happened', 'wasnt', 'talk', 'kissed', 'relationship']	The interplay of identity and relationship
1	7393	['alcohol tastes', 'alcohol taste', 'taste alcohol', 'alcohol like', 'make alcohol', 'drinking alcohol', 'like alcohol', 'drink alcohol', 'alcohol', 'drink like']	The diverse sensory perceptions of alcohol consumption
2	569	['parents weed', 'smoking weed', 'smoked weed', 'smoke weed', 'smokes weed', 'marijuana', 'weed', 'caught smoking', 'smoked', 'drugs']	Discussing marijuana within family dynamics and parental influence.
3	568	['buy cocaine', 'love cocaine', 'dont cocaine', 'making cocaine', 'cocaine dont', 'cocaine makes', 'cocaine cocaine', 'doing cocaine', 'cocaine', 'cocaine like']	Cocaine use: secretive activities and personal stories
4	548	['throw party', 'party alcohol', 'partying', 'school party', 'going party', 'party', 'house party', 'hosting party', 'party parents', 'birthday party']	Party culture: drugs, alcohol, and social interactions
5	496	['pens pencils', 'juuling', 'colored pencils', 'pencils', 'pen tube', 'pen', 'pencil case', 'gel pens', 'fountain pens', 'like pens']	The Risks and Consequences of Vaping and Nicotine Addiction Amongst Youth
6	422	['dad', 'father', 'dads', 'parents', 'mom', 'mother', 'mum', 'abusive', 'divorce', 'family']	Substance use and mental health
7	275	['classroom', 'student', 'teacher', 'expelled', 'students', 'th	Classroom discussions intersecting with

		grade', 'principal', 'grade', 'class', 'middle school']	substance use topics
8	265	['suicidal', 'suicide', 'depressed', 'depression', 'mental health', 'self', 'therapy', 'die', 'feel', 'kill']	Linking mental health struggles with substance use
9	264	['taking adderall', 'adderall', 'prescribed adderall', 'adhd meds', 'adderall like', 'took adderall', 'adderall time', 'diagnosed adhd', 'adhd', 'adderall prescription']	Personal accounts of Adderall usage and effects
10	220	['songs', 'indie rock', 'song listen', 'alternative rock', 'albums', 'music', 'playlists', 'playlist', 'hip hop', 'listening cigarettes']	Music as a backdrop to substance use discussions
11	185	['parents', 'family', 'mother', 'father', 'mom', 'dad', 'mum', 'sister', 'siblings', 'abusive']	Family involvement in substance use
12	129	['getting drunk', 'drunk went', 'got drunk', 'drunk friends', 'drank', 'drunk', 'drinking', 'drink', 'drinks', 'drunk time']	Social drinking experience
13	118	['addicted nicotine', 'nicotine addiction', 'quit vaping', 'quitting cigarettes', 'stop vaping', 'quit smoking', 'stop smoking', 'nicotine', 'addicted cigarettes', 'started vaping']	Struggles with nicotine addiction and quitting.
14	115	['drunk bored', 'bored drunk', 'talk drunk', 'im drunk', 'talk bored', 'im bored', 'drunk wanna', 'bored drank', 'drunk talk', 'bored im']	Social drink and ennui
15	91	['rap', 'gang gucci', 'gucci gang', 'niggas', 'bitch wanna', 'im nigga', 'make bitch', 'nigga', 'bitch', 'gang oofer']	Rap Culture and identify

Feature importance and correlation analysis

Figure 4 depict the coefficients measuring the importance of emotional features, as determined by the logistic regression model. The model's pseudo-R-squared value is approximately 0.074, and it has a current function value of 0.64. Coefficients with higher absolute values indicate stronger associations with substance use. The top five emotional value is relief (3.36), disgust (0.87), disapproval (0.72), amusement (0.46), and annoyance (0.36).

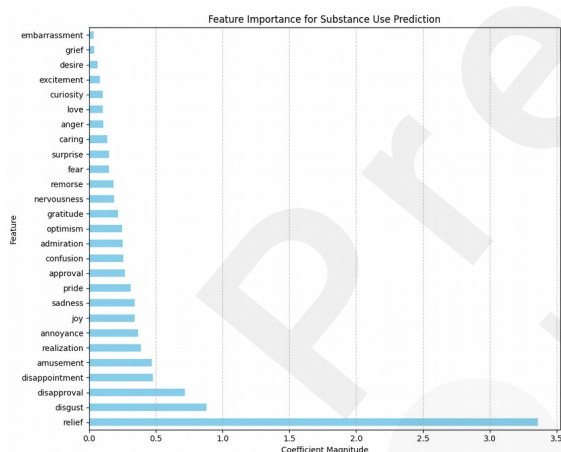


Figure 4 Emotional feature importance

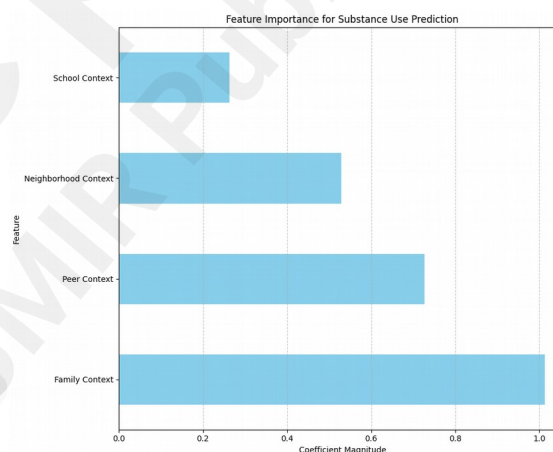


Figure 5 Contextual feature importance

Figure 5 shows the coefficients value for the contextual features. The logistic regression model successfully optimized, with a current function value of 0.65, was applied on the analysis. The family context has the largest coefficient of, indicating the strongest association. Peer context also shows a strong positive association, with a coefficient of 0.73. This is followed by the neighborhood context with a coefficient of 0.53, and then the school context with a coefficient of 0.26.

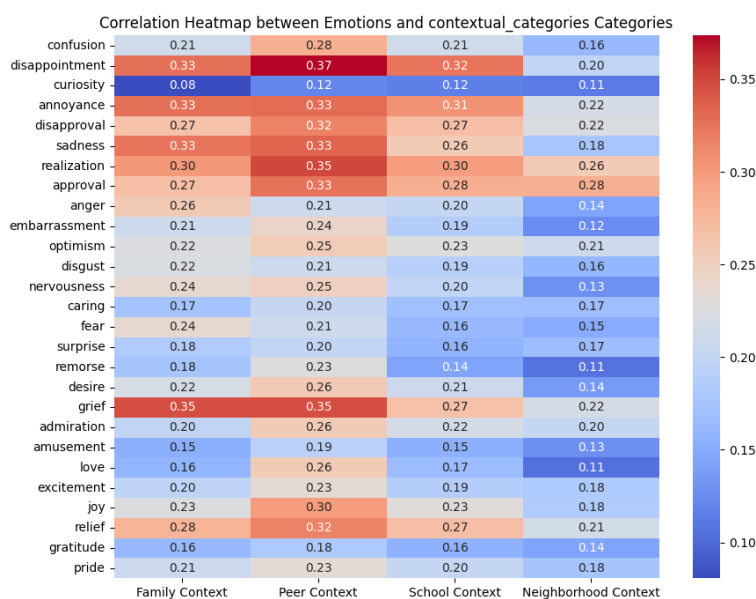


Figure6 Correlation heatmap between emotions and substance use and contextual categories

The correlation heatmap in Figure 6 shows the relationship between various emotion mean values and the four contextual categories. Among them, the peer context shows stronger positive correlations with most of emotion features than the other three contexts, particularly with emotions such as “disappointment” (0.37), “grief” (0.35), “realization” (0.35), “sadness” (0.33), and “disapproval” (0.32). In family context, there are both strong negative correlations with emotions like “grief” and “sadness”, and positive correlations with emotions such as “annoyance” (0.33), and “disappointment” (0.33). For the school context, “disappointment” (0.32) exhibits the highest positive correlation, while the neighborhood context features two positive correlations with “approval” (0.28) and “realization” (0.26).

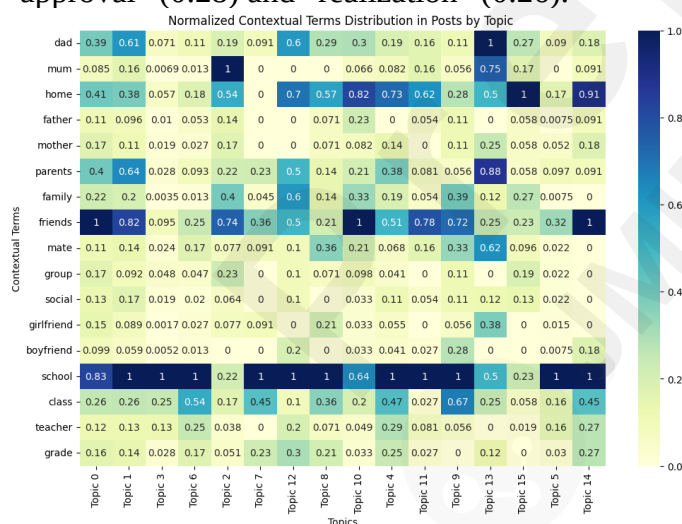


Figure 7 The correlation heatmap between top frequency contextual terms and topics

The heatmap Figure7 illustrates the intricate relationship between contextual terms related most social relationship and different topics within substance use discussion. The term "Dad" is common associated with familial roles and the setting of examples within the household, particularly when it comes to habits and behaviors around substance consumption and addiction. The term “mum” is often at the forefront of familial discussions about marijuana and smoking cessation.

The term ‘Home’ is deeply intertwined with the formative experiences related to substance use, particularly in the context of youth culture and family dynamic. The home environment plays a foundational role in shaping individuals’ encounters with substance use. It is where the earliest observations and imitations of social drinking occur, informed by the behavior of family members.

Additionally, music genres like rap, which can reflect substance use experiences, often finds resonate within the domestic sphere, impacting the youth's evolving self-concept. Furthermore, the home also serves as a battleground for addressing nicotine addiction, where family struggles and triumphs can motivate or deter such habits.

The term "friends" is highly relevant across topics, related to adolescence, relationships, music, and substance use indicates that peer relationships are a powerful influence on young people's substance-related behaviors.

The term "School" is prevalent term across various topics related to substance use, reflecting the pivotal role educational institutions play in the formative years of adolescence. As a hub for forming complex relationships, schools become a platform where decisions about substance use are influenced by peer interaction, academic pressure, and extracurricular activities.

Discussion

The analysis of substance use-related posts within the r/Teenagers subreddit community provides valuable insights into the prevalence and dynamics of discussions surrounding substance use among teenagers. It is noticeable that the keyword "drink" and "weed" are two most frequently mentioned substance terms within the data. This is consistent with the fact that teenagers often encounter new experiences during adolescence, and the desire for acceptance can make them susceptible to manipulation. They may find the pleasure derived from alcohol or weed to be unparalleled compared to other experiences they have had [30].

The distribution of substance use-related posts peaked in July 2023, indicating potential variations in posting behavior over time. Another peak occurred during June and July 2019, which aligns with the observation that summer offers more idle time, increased opportunities for social activities, outdoor events such as dance festival and parties, and other group-oriented social settings may lead to greater exposure to drugs or peers using drugs [31]. Summer can be a crucial time for parents, society, and psychology to pay more attention and develop strategies to educate about the consequence of substance abuse and how to handle peer pressure. Additionally, the distribution of substance uses posts remained lower compared to before (2019) and after (2022) the COVID-19 period. This decrease appears to be essentially explained by the fact that teenagers were less exposed to school, society, parties, and community activities through home-based learning [32].

Negative emotion has been considered particularly important in the development and maintenance of substance use disorder [33], as emotion during teenagers years tend to fluctuate rapidly due to the unique nature of this age group. Our finding reveals significant differences in emotional mean values between substance use posts and non-substance use posts. Specifically, emotions such as annoyance, approval, sadness, and disappointment had higher mean values in substance use posts, indicating the complexity and intensity of discussions related to substance use. Understanding these emotional symptoms may raise concerns regarding emotional disorders or other harmful behaviors in place of substance use or relapse to substance use [34].

The contextual analysis revealed that peer context and family context are quite popular. The school context was also very significant in 2019, but this factor decreased during and after COVID-19. This can be explained by the transition of schools to remote learning. However, a different pattern is observed with the family context, which remains higher compared to before the pandemic, replacing the school context in the second rank. Family environment exerted significant indirect effects on adolescent alcohol use through peer influence, self-efficacy, and stress, and parental expectations significantly moderated all structural paths [35]. Additionally, through BERTopic analysis, various topics related to substance use discussions were identified, including the interplay of identity and

relationships, sensory perceptions of alcohol consumption, and discussions about marijuana within family dynamics. These topics provide insights into the multifaceted nature of substance use discussions and the diverse range of issues addressed within the community. We need to pay more attention to enabling family members to communicate better with teenagers, helping them alleviate the stress brought about by school and peer.

The feature importance found that emotions such as relief, disgust, and disapproval showed strong associations with substance use, underscoring the emotional underpinnings of discussions surrounding substance use. The psychopharmacological properties of alcohol and other substances may provide some temporary relief, and, in addition, the belief that these substances will alleviate negative affective states may be powerful as well[36]. The correlation heatmap demonstrated the complex interplay between emotional features and contextual categories. Peer context showed stronger positive correlations with most emotional features, highlighting the significant role of peer influence in shaping emotional responses related to substance use. Peer influence is regarded as one of the strongest determinants of juvenile delinquency and particularly adolescent substance use. A commonly held view is that social pressure from friends to use drugs and alcohol is a major contributor to substance use[37]. Emotional detachment from parents and peer pressure impacts the likelihood of other illicit substance use across the high school years. Prevention programs may benefit by focusing on delaying the onset of first use of substance use and by preventing emotional detachment from parents and peer pressure[38].

The analysis of contextual terms such as "Dad," "Mum," "Home," "Friends," and "School" provided insights into the familial, social, and environmental factors influencing substance use discussions among teenagers. These terms reflect the diverse influences and contexts in which substance use behaviors are discussed and experienced. A father's approach to social drinking, can influence a teenager future behavior and attitudes towards alcohol consumption. Mothers can set the tone for open discussions about marijuana within the context of changing societal norms, and their experience. The effects of substance use are felt by whole family. The family context holds information about how substance use disorders develop, are maintained, and what can positively or negatively influence the treatment of the substance use disorder. Understanding the current developmental stage a family is in helps inform assessment of impairment and determination of appropriate interventions [39]. Parents' substance use may affect their ability to consistently prioritize the child's basic physical and emotional needs over their own need for substances. Friends often acts as a bridge between the private sphere of family influence and the broader cultural narratives found in music and social settings. The influence of friends is a potent force in solidifying or challenging substance use patterns developed. Family environment as well as peer groups has an important bearing on initiation and continuation of substance use[40]. Family members and friends were found to have a considerable influence not only on initiation but also important sources for money as well as the substance. Easy availability in the neighborhood was also an important correlate to continuation of substances. It is necessary to provide health promotion programs directed toward students and their families which encourage attitude shaping among school children towards self-confidence and adequacy.

Limitations and future research

While this study provides valuable insight into the substance use experience of teenagers within the online community, it is important to acknowledge certain inherent limitations in the methodology and models employed. Focusing only on the r/Teenagers subreddits restricts the generalizability of our finding to the broader teenage population. Moreover, due to the anonymous nature of Reddit

posts, we could not control for demographics factors such as age, gender, ethnicity, or country of origin. Future research should broaden its scope to include other social media platforms like Twitter and Facebook, which may provide access to demographic information. Conducting a comparative analysis across multiple social media platforms would offer a more comprehensive understanding of teenage substance use behaviors and experience.

Additionally, a study has developed a machine learning algorithm that predicts the age segment of Reddit users as adolescents or adults. By capturing linguistic patterns, posting behavior, and account details, this algorithm distinguishes between adolescent and adult age group[41]. Integrating such predictive models into future research could help overcome the challenge of identifying the age group of users posting on r/Teenagers, thereby enhancing the validity and reliability of our findings.

The contextual factors and topic examined emerged from the original submission posts. However, we are still far from being able to measure the risk factor may influence substance use in teenagers comprehensively. Future studies should incorporate comments to scale up the risk level. For instance, one study utilized a scale developed by clinical domain experts with 12 categories indicative of fentanyl misuse risk and then built machine learning classification models to identify fentanyl risk using Reddit posts and comments [42].

In summary, the posts provide snapshots of individuals' experiences and mental health states, external situational factors may influence the level of detail that can be inferred from the text. Moreover, the BERT model utilized for extracting emotion and topic features has its limitations in accuracy, leaving room for improvement. Future research could explore the combination of hand-labeling, thereby enhancing the overall effectiveness of analysis.

Conflicts of Interest

Disclose any personal financial interests related to the subject matters discussed in the manuscript here. For example, authors who are owners or employees of Internet companies that market the services described in the manuscript will be disclosed here. If none, indicate with "none declared".

References

1. Waterman EA, Lee KDM, Edwards KM. Longitudinal Associations of Binge Drinking with Interpersonal Violence Among Adolescents. *J Youth Adolesc.* 2019;48(7):1342-1352. doi:10.1007/s10964-019-01035-w
2. Degenhardt L, Stockings E, Patton G, Hall W, Lynskey M. The increasing global health priority of substance use in young people. *Lancet Psychiatry.* 2016;3 3:251-264. doi:10.1016/S2215-0366(15)00508-8
3. Gray K, Squeglia L. Research review: What have we learned about adolescent substance use? *J Child Psychol Psychiatry.* 2018;59. doi:10.1111/jcpp.12783
4. Wolf JP, Lipperman-Kreda S, Bersamin M. "it just depends on the environment": Patterns and decisions of substance use and co-use by adolescents. *J Child Adolesc Subst Abuse.* 2019;28(3):143-149. doi:10.1080/1067828x.2019.1637316
5. Petraitis J, Flay BR, Miller TQ. Reviewing theories of adolescent substance use: Organizing pieces in the puzzle. *Psychol Bull.* 1995;117(1):67-86. doi:10.1037/0033-2909.117.1.67
6. *Complete Guide to Substance Use + Mental Health.* Child Mind Institute; 2023.
7. Mason M, Mennis J, Russell M, Moore M, Brown A. Adolescent depression and substance use: The protective role of prosocial peer behavior. *J Abnorm Child Psychol.* 2019;47(6):1065-1074. doi:10.1007/s10802-018-0501-z
8. Chen HJ, Balan S, Price RK. Association of contextual factors with drug use and binge

drinking among white, native American, and mixed-race adolescents in the general population. *J Youth Adolesc.* 2012;41(11):1426-1441. doi:10.1007/s10964-012-9789-0

9. Gallupe O, Bouchard M. Adolescent parties and substance use: A situational approach to peer influence. *J Crim Justice.* 2013;41(3):162-171. doi:10.1016/j.jcrimjus.2013.01.002

10. Loke AY, Mak Y wah. Family process and peer influences on substance use by adolescents. *Int J Environ Res Public Health.* 2013;10(9):3868-3885. doi:10.3390/ijerph10093868

11. Muchiri BW, dos Santos MML. Family management risk and protective factors for adolescent substance use in South Africa. *Subst Abuse Treat Prev Policy.* 2018;13(1):24. doi:10.1186/s13011-018-0163-4

12. Eisenberg ME, Toumbourou JW, Catalano RF, Hemphill SA. Social Norms in the Development of Adolescent Substance Use: A Longitudinal Analysis of the International Youth Development Study. *J Youth Adolesc.* 2014;43(9):1486-1497. doi:10.1007/s10964-014-0111-1

13. Lynn Mulvey K, Boswell C, Zheng J. Causes and Consequences of Social Exclusion and Peer Rejection Among Children and Adolescents. *Rep Emot Behav Disord Youth.* 2017;17(3):71-75.

14. Egan KL, Suerken C, Debinski B, et al. More than just Alcohol: Marijuana and Illicit Drug Use at Parties Attended by 15–20 Year Olds. *Subst Use Misuse.* 2019;54(2):297-306. doi:10.1080/10826084.2018.1517798

15. Bisset S, Markham WA, Aveyard P. School culture as an influencing factor on youth substance use. *J Epidemiol Community Health.* 2007;61(6):485-490. doi:10.1136/jech.2006.048157

16. Bisset S, Markham WA, Aveyard P. School culture as an influencing factor on youth substance use. *J Epidemiol Community Health.* 2007;61(6):485-490. doi:10.1136/jech.2006.048157

17. Piko BF, Kovács E. Do parents and school matter? Protective factors for adolescent substance use. *Addict Behav.* 2010;35(1):53-56. doi:10.1016/j.addbeh.2009.08.004

18. Allison K, Crawford I, Leone P, et al. Adolescent substance use: Preliminary examinations of school and neighborhood context. *Am J Community Psychol.* 1999;27:111-141. doi:10.1023/A:1022879500217

19. Fagan A, Wright E, Pinchevsky GM. Exposure to violence, substance use, and neighborhood context. *Soc Sci Res.* 2015;49:314-326. doi:10.1016/j.ssresearch.2014.08.015

20. Deas N, Kowalski R, Finnell S, et al. I just want to matter: Examining the role of anti-mattering in online suicide support communities using natural language processing. *Comput Hum Behav.* 2023;139:107499. doi:10.1016/j.chb.2022.107499

21. Reddit comments/submissions 2005-06 to 2022-12 : reddit: free download, borrow, and streaming. Internet Archive. Accessed February 22, 2024. https://archive.org/details/pushshift_reddit_200506_to_202212

22. Iversen GR. *Contextual Analysis*. SAGE; 1991.

23. Demszky D, Movshovitz-Attias D, Ko J, Cowen A, Nemade G, Ravi S. GoEmotions: a dataset of fine-grained emotions. Published online June 2, 2020. doi:10.48550/arXiv.2005.00547

24. Vayansky I, Kumar SAP. A review of topic modeling methods. *Inf Syst.* 2020;94:101582. doi:10.1016/j.is.2020.101582

25. Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure. Published online March 11, 2022. doi:10.48550/arXiv.2203.05794

26. Grootendorst M. BERTopic: What is so special about v0.16? Maarten Grootendorst. Published December 19, 2023. Accessed February 21, 2024. <https://www.maartengrootendorst.com/blog/bertopic/>

27. Understanding feature importance in machine learning | built in. Accessed March 9, 2024. <https://builtin.com/data-science/feature-importance>
28. Pandas.DataFrame.corr — pandas 2.2.1 documentation. Accessed March 9, 2024. <https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.corr.html>
29. Sharma P, Li Y. *Self-Supervised Contextual Keyword and Keyphrase Retrieval with Self- Labelling*. MATHEMATICS & COMPUTER SCIENCE; 2019. doi:10.20944/preprints201908.0073.v1
30. Chronister DK, Program KHTT, Program DKC and KHTT. The reasons why teens like weed and alcohol might not shock you. Teen Treatment Center in Los Angeles | Key Healthcare. Published July 29, 2021. Accessed March 11, 2024. <https://keyhealthcare.com/the-reasons-why-teens-like-weed-and-alcohol-might-not-shock-you/>
31. Summertime associated with highest risk for teen drug use. Narconon International. Accessed March 11, 2024. <https://www.narconon.org/blog/summertime-associated-with-highest-risk-for-teen-drug-use.html>
32. Bourdige C, Teissedre F, Morel F, Flaudias V, Izaute M, Brousse G. Lockdown impact on stress, coping strategies, and substance use in teenagers. *Front Psychiatry*. 2022;12:790704. doi:10.3389/fpsyt.2021.790704
33. Kang D, Fairbairn CE, Ariss TA. A meta-analysis of the effect of substance use interventions on emotion outcomes. *J Consult Clin Psychol*. 2019;87(12):1106-1123. doi:10.1037/ccp0000450
34. Yang G, King SG, Lin HM, Goldstein RZ. Emotional expression on social media support forums for substance cessation: Observational study of text-based reddit posts. *J Med Internet Res*. 2023;25:e45267.
35. Nash SG, McQueen A, Bray JH. Pathways to adolescent alcohol use: Family environment, peer influence, and parental expectations. *J Adolesc Health*. 2005;37(1):19-28. doi:10.1016/j.jadohealth.2004.06.004
36. Trull T. Borderline personality disorder and substance use disorders a review and integration. *Clin Psychol Rev*. 2000;20(2):235-253. doi:10.1016/S0272-7358(99)00028-8
37. Reed MD, Rountree PW. Peer pressure and adolescent substance use. *J Quant Criminol*. 1997;13(2):143-180. doi:10.1007/BF02221306
38. Gallegos MI, Zaring-Hinkle B, Wang N, Bray JH. Detachment, peer pressure, and age of first substance use as gateways to later substance use. *Drug Alcohol Depend*. 2021;218:108352. doi:10.1016/j.drugalcdep.2020.108352
39. Lander L, Howsare J, Byrne M. The impact of substance use disorders on families and children: From theory to practice. *Soc Work Public Health*. 2013;28(3-4):194-205. doi:10.1080/19371918.2013.759005
40. Tsering D, Pal R. Role of family and peers in initiation and continuation of substance use. *Indian J Psychol Med*. 2009;31(1):30-34. doi:10.4103/0253-7176.53312
41. Chew R, Kery C, Baum L, Bukowski T, Kim A, Navarro M. Predicting age groups of reddit users based on posting behavior and metadata: Classification model development and validation. *JMIR Public Health Surveill*. 2021;7(3):e25807. doi:10.2196/25807
42. Garg S, Taylor J, El Sherief M, et al. Detecting risk level in individuals misusing fentanyl utilizing posts from an online community on reddit. *Internet Interv*. 2021;26:100467. doi:10.1016/j.invent.2021.100467