Differential Analysis of Age, Gender, Race, Sentiment, and Emotion in Substance Use Discourse on Twitter during the COVID-19 Pandemic: An NLP Approach

Julina Maharjan, Ruoming Jin, Jennifer King, Jianfeng Zhu, Deric Kenne

Submitted to: Journal of Medical Internet Research
on: October 08, 2024

Table of Contents

Original Manuscript..... 5

Differential Analysis of Age, Gender, Race, Sentiment, and Emotion in Substance Use Discourse on Twitter during the COVID-19 Pandemic: An NLP Approach

Julina Maharjan[1] PhD; Ruoming Jin[1] Prof Dr; Jennifer King[2] Prof Dr; Jianfeng Zhu[1] PhD; Deric Kenne[2] Prof Dr

[1]Department of Computer Science Kent State University Kent US
[2]Department of Public Health Kent State University Kent US

Corresponding Author:
Julina Maharjan PhD
Department of Computer Science
Kent State University
800 E. Summit St.
Kent
US

Abstract

Background: User Demographics are often hidden in social media data due to privacy concerns. However, demographic information on Substance Use can provide valuable insights, allowing Public Health policymakers to focus on specific cohorts and develop efficient prevention strategies, especially during global crises like COVID-19.

Objective: Our study aims to analyze Substance Use trends in User level across different demographic dimensions; such as Age, Gender and Race/Ethnicity, focusing on COVID-19 pandemic. The study also establishes a baseline for substance use trends using social media data.

Methods: The study is carried out in large scale Twitter data in the English language over a 3 year period; 2019, 2020 and 2021, which comprises 1.05 billions of posts. Following preprocessing, the substance use posts were identified using our custom trained deep learning model (RoBERTa) that resulted in identification of 9 million Substance Use posts. Then, demographic attributes like User Type, Age, Gender, Race/Ethnicity, and Sentiment types, and emotions associated with each post were extracted via a collection of natural language processing modules. Finally, various qualitative analyses were performed to get the insight of user behaviors based on the demographics.

Results: The highest level of usership in SU discussions was observed in 2020, with increases of 22.18% compared to 2019 and 25.24% compared to 2021. Throughout the study period, Male and Teenagers increasingly dominated the Substance Use discussions in all substances. During the pandemic, Prescription Medication among Female usership was observed high compared to other substances. Additionally, Alcohol usership increased by 80% within two weeks after the Global Pandemic declaration in 2020.

Conclusions: Our study presents a large-scale, fine-grained analysis of Substance Use on social media data by age, gender and race/ethnicity before, during, and after COVID-19 pandemic. Overall, our analysis from social media data provides a new baseline study for substance usage that can help in prevention of substance use in an efficient manner.

(JMIR Preprints 08/10/2024:67333)
DOI: https://doi.org/10.2196/preprints.67333

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

Journal of Medical Internet Research

Differential Analysis of Age, Gender, Race, Sentiment, and Emotion in Substance Use Discourse on Twitter during the COVID-19 Pandemic: An NLP Approach

Abstract

Background:
User Demographics are often hidden in social media data due to privacy concerns. However, demographic information on Substance Use can provide valuable insights, allowing Public Health policymakers to focus on specific cohorts and develop efficient prevention strategies, especially during global crises like COVID-19.
Objective: Our study aims to analyze Substance Use trends in User level across different demographic dimensions; such as Age, Gender and Race/Ethnicity, focusing on COVID-19 pandemic. The study also establishes a baseline for substance use trends using social media data.
Methods: The study is carried out in large scale Twitter data in the English language over a 3 year period; 2019, 2020 and 2021, which comprises 1.05 billions of posts. Following preprocessing, the substance use posts were identified using our custom trained deep learning model (RoBERTa) that resulted in identification of 9 million Substance Use posts. Then, demographic attributes like User Type, Age, Gender, Race/Ethnicity, and Sentiment types, and emotions associated with each post were extracted via a collection of natural language processing modules. Finally, various qualitative analyses were performed to get the insight of user behaviors based on the demographics.
Results: The highest level of usership in SU discussions was observed in 2020, with increases of 22.18% compared to 2019 and 25.24% compared to 2021. Throughout the study period, Male and Teenagers increasingly dominated the Substance Use discussions in all substances. During the pandemic, Prescription Medication among Female usership was observed high compared to other substances. Additionally, Alcohol usership increased by 80% within two weeks after the Global Pandemic declaration in 2020.
Conclusions: Our study presents a large-scale, fine-grained analysis of Substance Use on social media data by age, gender and race/ethnicity before, during, and after COVID-19 pandemic. Overall, our analysis from social media data provides a new baseline study for substance usage that can help in prevention of substance use in an efficient manner.

Keywords: substance use; social media; deep learning; NLP; COVID-19; age; gender; race; sentiment; emotion;

Introduction

Substance Use (SU) prevalence varies across demographics such as age, gender, and race/ethnicity. During the COVID-19 pandemic, these differences became more pronounced. The pandemic not only increased global substance use, with overdose deaths rising by 29.4% [1], but also exacerbated societal and racial inequalities [5, 6] , and significantly impacted mental health [7 - 10]. As people often turn to substances as a coping mechanism during crises [11, 12], the pandemic likely led to increased substance use [13], particularly among vulnerable populations [14]. Investigating how these trends shifted across different demographic groups during the pandemic is crucial for understanding public health challenges and developing targeted interventions.

Background

Gender, Age, and Racial Disparities in Substance Use

The annual drug statistics as per National Center for Drug Abuse Statistics (NCDAS) [4] shows that more men are likely to consume illicit drugs. In 2020, 22% of males and 17% of females used illegal drugs or misused prescription drugs within the last year, and was observed highest among persons between the ages of 18-25 at 39% compared to persons aged 26-29, at 34% [3]. Subsequently, racial and ethnic disparities have always been prevalent in the history of drug usage. For instance, White were more likely to misuse prescription drugs while other races were more likely to use other illicit drugs [15]. Similarly, Opioid overdose death rates were higher in Black [16]. And the disparities by race/ethnicity are also found to be varied with age, such that for most Substance Use Disorders (SUDs), estimated prevalence were higher for White participants at younger ages and Black participants at older ages [17].

Importance of Studying Substance Use During the COVID-19 Pandemic

Given the pre-existing disparities in Substance Use, the COVID-19 pandemic likely exacerbated these trends. According to CDC [1], COVID-19 mortality rates from January 1, 2020, to May 31, 2024, varied significantly by age, gender, and race/ethnicity. Non-Hispanic Whites accounted for 67% of deaths, individuals aged 75 and older represented approximately 54% of deaths, and males comprised 54% of the mortality rate. Simultaneously, the COVID-19 pandemic brought significant social and economic changes, disproportionately affecting minority and underprivileged populations [17, 19]. The rapid spread of the virus overwhelmed healthcare services, leading to lower priority for treatment for people of color and economically disadvantaged individuals [18, 20]. This discrimination exacerbated mental health issues [7], also highlighted by the CDC [1], which noted disparities in mental health and substance misuse among racial and ethnic minority populations due to unequal access to care, psychosocial stress, and social determinants of health. Given the disparities in COVID-19 mortality rates by age, gender, and race/ethnicity, and the social and economic challenges exacerbated by the pandemic, studying substance use trends across different demographic groups requires high attention. The disproportionate impact on minority and underprivileged populations highlights the need to understand how these factors influenced substance use, which will aid in developing targeted public health strategies to address the specific needs of affected populations.

Related Study

The study of substance use prevalence across demographics has predominantly relied on survey-based research conducted by national agencies such as Substance Abuse and Mental Health Service Administration (SAMHSA) [3] and National Institute of Drug Abuse (NIDA) [2]. For example, the National Survey on Drug Use and Health (NSDUH) [4], administered by SAMHSA, provides comprehensive data on substance use and mental health issues among the U.S. population aged 12 and older. Similarly, the Monitoring the Future (MTF) [21] survey, funded by NIDA, focuses on substance use patterns among youth by surveying middle and high school students (grades 8, 10, and 12). Both surveys provide detailed reports on the use of illicit and non-illicit drugs, disaggregated by age, gender, and race/ethnicity at a national level. In addition to these national surveys, various individual studies [17, 22] also have explored substance use disparities across demographics such as age, gender, and race/ethnicity.

While these surveys offer valuable insights, their scope is often limited by the diversity of true populations and duration of studied period. Traditional survey methods often rely on self-reported data, which can be affected by social desirability bias and recall errors. Additionally,

surveys are typically conducted annually or biennially, providing only periodic snapshots of substance use trends. On the other hand, COVID-19 brought additional challenges in data collection. SAMHSA 2020, for instance, was able to collect only first and fourth quarter data due to in-person restriction caused by COVID-19 pandemic [3].

In contrast, social media data addresses many of these limitations. Social media platforms capture real-time, user-generated content that often reflects more authentic behaviors and sentiments. In addition to this, researchers have also shown the prevalence of substance use discussion in social media [23-30] possibly due to its anonymity feature. Likewise, the continuous stream of data allows researchers [31, 24, 26, 38, 32] to monitor trends as they evolve, providing insights that are not possible with traditional survey methods. Additionally, the vast amount of data available on social media enables a more detailed analysis across a large population [32], including those that might be underrepresented in surveys [30, 33].

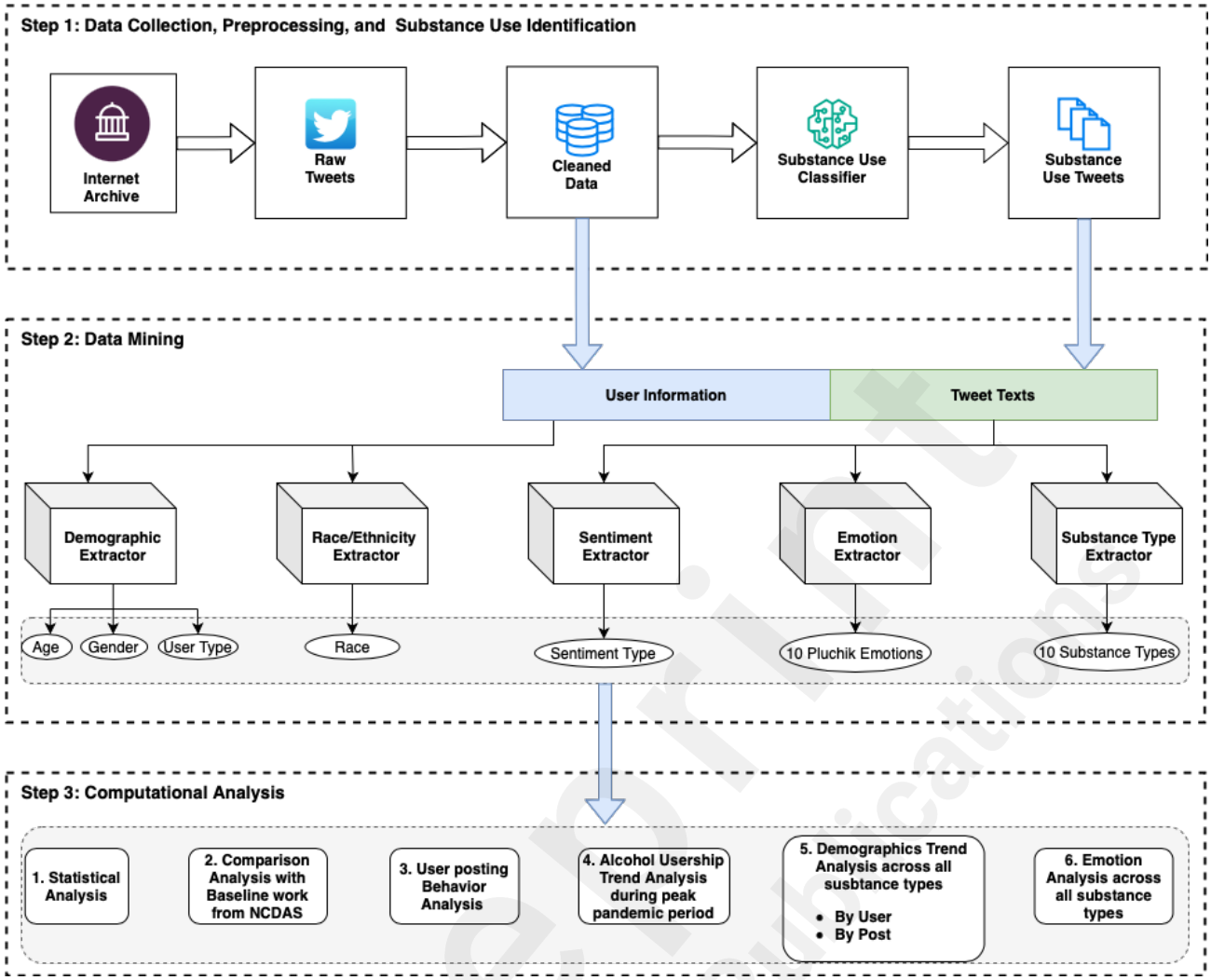
Despite the extensive research on substance use trends, there remains a gap in understanding how these trends vary across different demographic groups, especially in the context of the COVID-19 pandemic. Existing studies have primarily relied on less diverse survey data or real time data in a small span of period, often overlooking the dynamic and nuanced shifts in substance use behavior during global crises. Our study aims to address this gap by leveraging a large-scale social media dataset to provide a more granular and continuous analysis of substance use trends across diverse demographics before, during, and after the pandemic. The following research questions are designed to explore these trends in detail, offering insights into how age, gender, race/ethnicity, and emotional factors have influenced substance use patterns during the study period.

1. What are the statistical distributions of overall SU posts and their users, categorized by key demographic variables in pre-pandemic, pandemic and post-pandemic periods?
2. To what extent do the substance use patterns and demographic distributions observed in Twitter discourse from 2019 to 2021 correspond with or differ from the baseline trends reported by the National Center for Drug Abuse Statistics (NCDAS) and/or other research?
3. What are the temporal trends in SU posts across different substance types throughout the study period, and how does the frequency of user posting behavior vary over time for each substance type?
4. How did the number of individuals using Alcohol change in the immediate aftermath of the pandemic declaration compared to those using other substances, and what short-term trends in user behavior across different age groups, genders, races and sentiments during this period?
5. What is the trend in usership across different demographics in each substance type?
6. What emotional expressions are prevalent across all substance types?

Data and Methods

The demographic analysis conducted in this study is based on the substance use posts identified in our prior research [35]. Consequently, we focus exclusively on the data extraction processes pertinent to this investigation. As illustrated in Figure 1, Step 1 encompasses the data collection, preprocessing, and substance use identification modules that were previously detailed in [35]. In this research, we commence at Step 2, referred to as data mining, wherein we integrate multiple analytical modules to extract demographic information from user profiles and to derive emotional and sentiment classifications from the Twitter posts. Finally, step 3 involves computational analysis of the extracted data.

Figure 1. Methodology of study design



Data Collection, Preprocessing, and Substance Use Identification

The research presented in this paper builds upon the substance use posts identified in our prior work [35], which successfully documented substance use posts from January 2019 to December 2021. Specifically, we collected raw Twitter data from the Internet Archive [34] and underwent multiple rounds of preprocessing to clean the data (detailed procedures can be found in [35]). The cleaned tweets were then analyzed using a deep learning-based substance use classifier, a pivotal element of our earlier research. This classifier was developed utilizing a state-of-the-art model, RoBERTa [41], in conjunction with various techniques such as transfer learning and a Human-in-the-Loop approach [42] to enhance its performance, achieving an accuracy rate of 80%. The entire workflow for data collection, preprocessing, and substance use identification is depicted in Step 1 of Figure 1. This comprehensive phase concluded with a total of 2.8 million, 3.5 million, and 2.5 million substance use posts identified for the years 2019, 2020, and 2021, respectively.

Data Mining

Data mining constitutes a critical initial phase for conducting this research. This study utilized two distinct types of tweet data (user information and tweet texts) to extract essential variables, including age, gender, user type, race/ethnicity, sentiment, emotion, and substance

type. Initially, we aggregated posts based on unique user IDs to facilitate an analysis from the user perspective. Subsequently, we employed five different analytical modules: two focused on demographic variables (M3-Inference [38] and Ethicolr [41]), and three targeted at other relevant variables (VADER [39], SpanEmo [40], and a Substance Type Extractor developed in prior research [35]). Specifically, M3-Inference [36] and Ethicolr [41] were utilized to extract demographic information such as age, gender, user type, and race/ethnicity. In parallel, VADER [39], SpanEmo [40], and the Substance Type Extractor [35] were employed to extract sentiment, emotional content, and substance type, respectively. In the following section, we provide detailed descriptions on each of the modules.

Demographic (Age, Gender, UserType) Extraction using M3Inference

In this study, we implemented the M3-Inference model [38] to extract demographic information, specifically age group, gender, and user type, from Twitter accounts. M3-Inference is an open-source Python implementation of a multimodal deep learning system, trained on extensive datasets, including Twitter, IMDB, and Wikipedia. The model's architecture enables it to simultaneously predict three key demographic attributes: Multimodal capabilities, which allow processing of both image and text features (we only utilized text features to perform our work); Multilingual support, which includes 32 languages; and Multi-attribute prediction, which facilitates simultaneous forecasting of age, gender, and user type.

In terms of classification, the model treats gender (female or male) and user type (human or organization) as binary classification tasks, while age is categorized into four distinct groups: ≤18, 19-29, 30-39, and ≥40 years. For our analysis, we utilized a text-only pipeline to derive demographic predictions. This pipeline involved generating character-based embeddings for each textual input (username, screen name, and biography) and passing them through a two-layer bidirectional character-level Long Short-Term Memory (LSTM) network.

To validate the M3-Inference model's efficacy in predicting demographic attributes, we collected profile information from 50 known Twitter users (as detailed in Multimedia Appendix Table 1). For the age classification, we combined the outputs from the 19-29, 30-39, and ≥40 sub-groups into a single non-teenager category, thereby reformulating the age prediction as a binary classification task. The model's performance metrics on the collected validation data indicated an Accuracy of 99.05% for user type, 95% for gender, and 89% for age classification, with corresponding F1-scores of 0.98, 0.94, and 0.73, respectively.

Race/Ethnicity Extraction using Ethnicolr

To infer the racial and ethnic backgrounds of individuals from names, we employed the Ethnicolr Python library [41]. This tool leverages several models based on different datasets, including U.S. Census data, Wikipedia entries, and Florida voter registration records, to predict the likelihood of an individual's race and ethnicity. The model has three models depending upon the type of dataset it is trained on. In our case, we used the Census Last Name Model which was trained on U.S Census data [43] from the years 2000 and 2010. This model estimates the percentage likelihood that an individual belongs to four main racial and ethnic categories such as White, Black, Asian/Pacific Islander (API), or Hispanic. The predictions are appended as additional columns in the dataset, providing a probabilistic breakdown of racial composition. We verified the model on our sample data where the model achieved 90% accuracy. The sample predicted data is presented in the Multimedia Appendix Table 2. However, in our study, we were able to extract the race information for only those posts which had the firstname and lastname present in the tweet posts, otherwise the identification was not accomplished. Hence, around 65% of the total users were only identified as explained in the

result section in Table 1.

Sentiment Extraction using VADER

VADER (Valence Aware Dictionary and sEntiment Reasoner) [39] is a fully open source sentiment analysis tool designed specifically for social media text. It combines a lexicon-based approach with contextual rules to determine the sentiment of text as positive, negative, or neutral. VADER's lexicon assigns sentiment scores to words on a scale from -4 (very negative) to +4 (very positive). Contextual adjustments are made through several mechanisms; punctuation, such as exclamation points, can amplify sentiments; capitalization highlights intensity, with all-caps being more emphatic; degree modifiers, like "very," strengthen sentiment; and conjunctions, such as "but," can alter sentiment direction. The tool calculates a compound score, ranging from -1 (very negative) to +1 (very positive), by summing these adjusted scores. This method enables VADER to effectively capture both explicit and nuanced emotional expressions, providing a quick and reliable measure of overall sentiment in large volumes of text. The predicted sentiments for sample tweets is presented in Multimedia Appendix Table 3.

Emotion Extraction using SpanEmo

SpanEmo [40], is a deep learning based multi-label emotion recognition model. It operates by analyzing segments of text; spans, and classifying each span according to the emotions it conveys. The keywords associated with each emotion class are presented in Multimedia Appendix Table 4. This is particularly useful in complex texts where different parts may express different emotions. The tool uses NLP techniques to understand the context and semantic meanings of words and phrases, which allows it to accurately detect emotions even in nuanced or mixed-emotional content. The model is based on the BERT encoder that takes |C| i.e 10, number of emotion classes and a sequence 's', as inputs formatted with standard tokens; start_of_token [CLS] and separator_token [SEP] as [CLS] + [C] + [SEP] + s. The encoding of emotion classed in the input makes the model learn association between the emotion classes and the words in the input sentence, which is why it outperforms existing emotion classifiers. The model outputs 10 multi-emotion classes; namely, anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise, and trust. Before utilizing this module, we finetune this model on the SemEval-2018 multilabel emotion classification data set [43] and achieved F1-Micro score of 0.70. The predicted emotions for sample tweets is presented in Multimedia Appendix Table 5.

Substance Type Identification

The substance type identification module is also based on our previous work [35]. In earlier work [35], we considered the ten primary substance types categorized together based on their pharmacological and behavioral effects, and used the list of keywords from NIDA [2] to formulate keyword based identification. The ten types of substances are Tobacco, Alcohol, Cannabinoids, Opioids, Stimulants, Club Drugs, Hallucinogens, Dissociative Drugs, Prescription Medications, and Other Compounds.

Computational Analysis

In this study, we employed two primary statistical techniques: trend and comparison analysis, along with sentiment and emotion analysis.

Trend and Comparison Analysis

To explore temporal patterns in substance use discussions, we conducted a trend analysis, examining the frequency of posts over time. This allowed us to compare substance use trends before, during, and after the pandemic. We further performed comparative analysis to assess differences in substance use discussions across demographic categories, including age, gender, and race, identifying key disparities and dominant trends.

Sentiment and Emotion Analysis:

We applied the VADER model to perform sentiment analysis, classifying the overall tone of posts (positive, negative, or neutral) related to substance use. Additionally, the SpanEmo model was used for emotion detection, allowing us to identify and categorize emotional expressions (e.g., joy, anger, sadness) linked to specific substances.

These methods provided insight into both the temporal dynamics of substance use discussions and the emotional context in which they occurred.

Ethical Considerations

To ensure the privacy and confidentiality of individuals whose data were analyzed, all study data underwent a rigorous de-identification process before analysis. The data for this study were sourced from publicly available platforms [34], containing no identifiable personal information. Additionally, the sample posts were preprocessed to transform them into tokens, effectively obscuring any details that could reveal users' identities. Consequently, the dataset used in this study is completely free of personal identifiers, such as author names or any other private information.

Results

In this study, we present a fine grained demographic analysis of substance usage in Twitter discourse from dual perspective; by post and by user. In total, after pre-processing and Substance Use identification, our final data set included 2,799,726, 3,502,171, 2,553,235 posts, and 2,131,457, 2,604,123, 2,553,235 users in 2019, 2020, and 2021 respectively. In the following sections, at first, we present a substantial summary of substance use trends across all demographic dimensions; User Type, Gender Type, Age Group, Race/Ethnicity type, and including Sentiment Type. Then, we compare our result with survey based baseline research from NCDAS Further, we analyze the usership trends on different substances, where Alcohol users were found to be the prime users during Peak Pandemic (March 2020 to June 2020). Hence, we perform a detailed analysis on Alcohol users/posts during this time. In addition to this, we present the trends of all substance users across five demographics for each substance type. Lastly, we present radar plots to understand the associated emotions with each substance type.

Question 1: What are the statistical distributions of overall SU posts and their users, categorized by key demographic variables in pre-pandemic, pandemic and post-pandemic periods?

Our key findings from the statistical analysis are presented in Multimedia Appendix Table 6, which summarizes the distribution of identified substance use (SU) posts by both posts and users, further segmented by various categories including user type, gender, age group, sentiment, and race/ethnicity. As indicated in source [46], the Twitter user base has been

expanding annually, with increases of 11.1% in 2020 and 4.25% in 2021. This growth is contextualized in Figure 2, where we illustrate the trends in the Twitter user base and substance use in 2020, comparing both posts and users to pre-pandemic and post-pandemic years. Notably, despite the increase in Twitter users, the marginal decline in substance use posts and users in 2021 implies that substance usage was significantly higher in 2020.

Figure 2. Overview of trends in Twitter Users and Substance Users/Posts

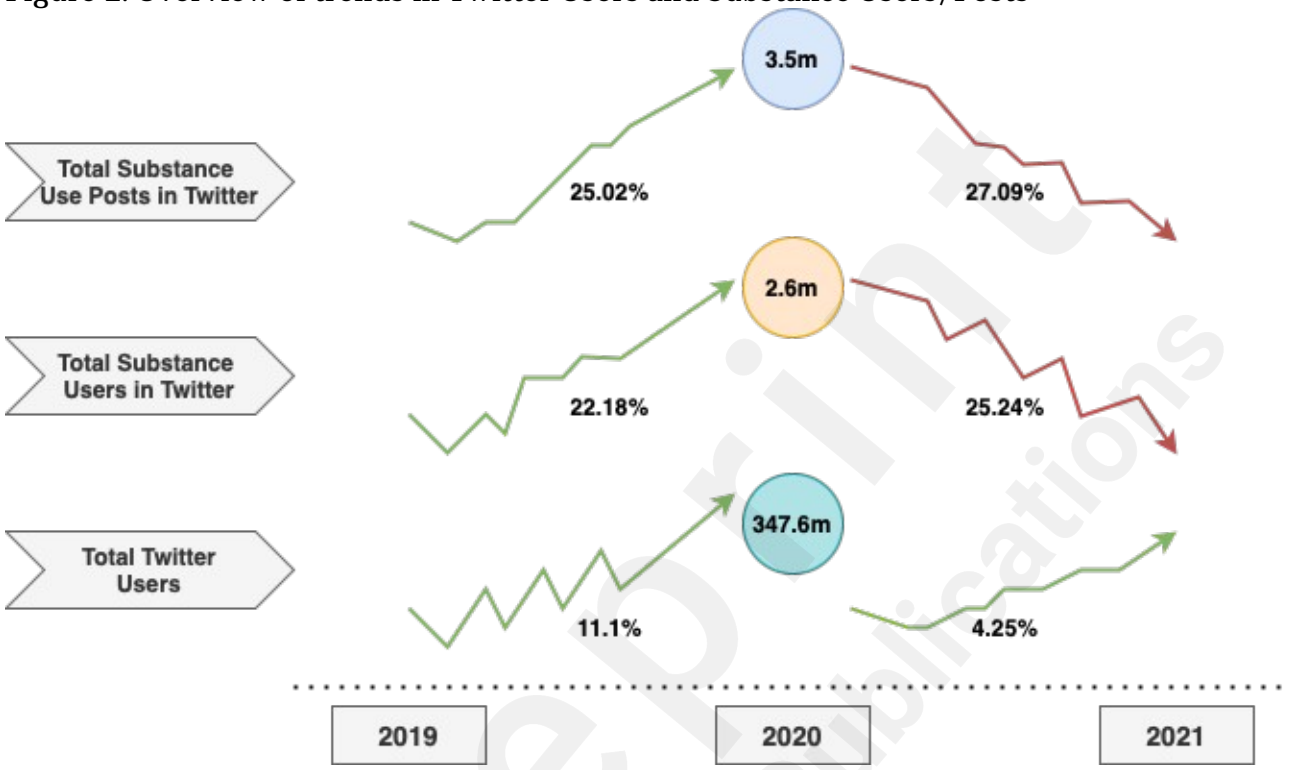


Figure 3. Trends in SU usership by gender, age group, and race/ethnicity

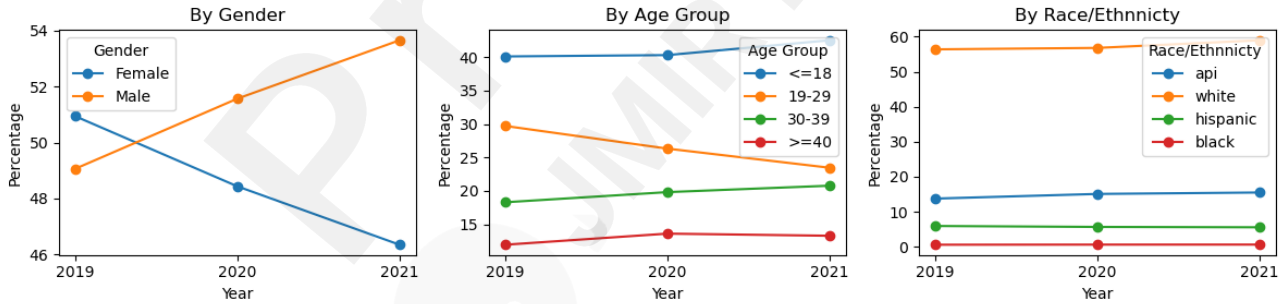
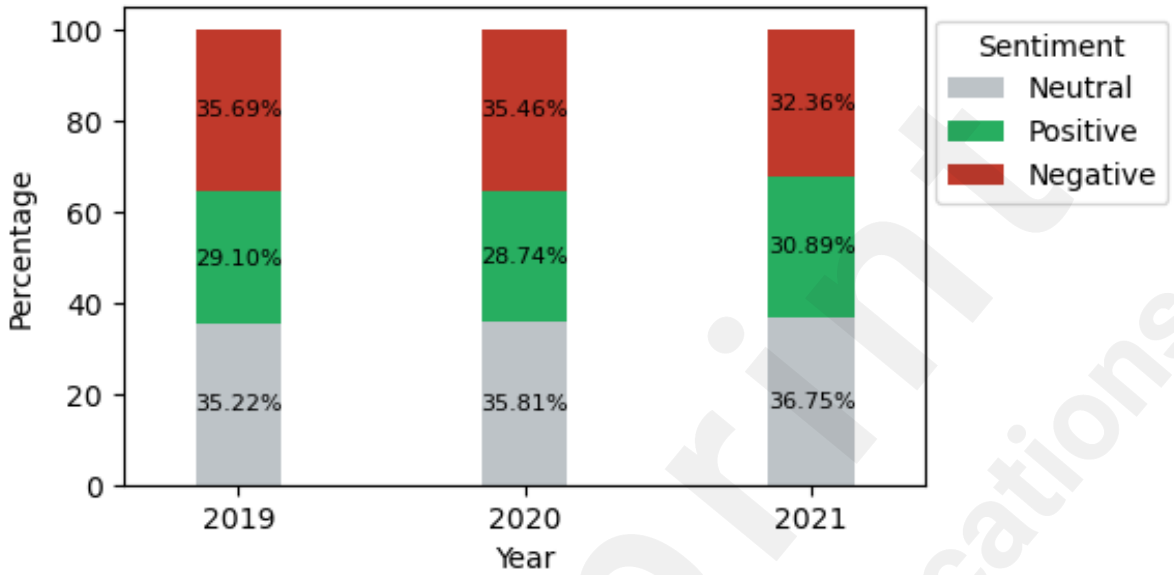


Figure 3 presents the line plots depicting substance use among the Twitter user base, categorized by gender, age group, and race/ethnicity. Statistics from twitter users by gender [46] revealed that male users consistently outnumber female users on Twitter, with a distribution of 68% male and 32% female in 2020. In contrast, our analysis indicates that among substance users in 2020, 52% were male and 48% were female. This suggests that, despite a smaller female user base, female substance users represent a significant proportion of the overall female demographic on Twitter. Furthermore, the data, as shown in Figure 3 (by gender), indicates a growth trend among female substance users from 2019 to 2021, whereas male substance users exhibit a declining trend during the same period.

Similarly, according to statistics from Twitter users [46], the highest levels of user engagement is found in the age group of 18-35. Moreover, our analysis also revealed a similar trend in substance users as shown in Figure 3 (by age group), where a greater number of younger users

are identified as substance users. Notably, our analysis indicates an increasing trend among teenagers (<=18) while showing a decline in substance use among the 19-29 age group. In addition to this, our findings in racial groups, as shown in Figure 3 (by race/ethnicity), reveal that White users consistently outnumber users from other racial groups across all study periods, likely due to the predominance of English language tweets in our analysis.

Figure 4 Sentiment distribution in SU posts from 2019 through 2021

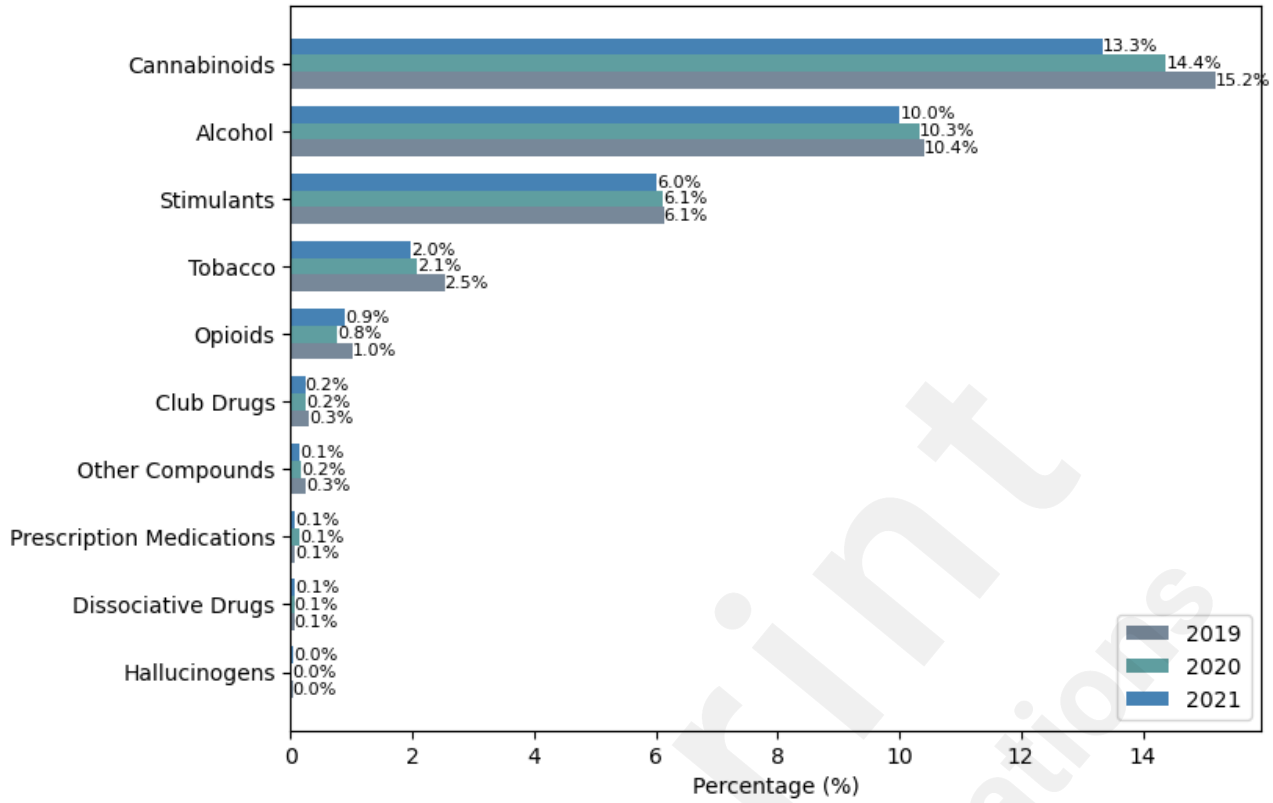


In addition to this, the sentiment distribution of substance use posts revealed that substance use posts during 2019 and 2020 were highly associated with negative comments compared to 2021, as shown in Figure 4. Nevertheless, the inclining trend in positive and neutral comments in post pandemic suggests that 2020 had relatively high negative factors on substance use.

Question 2: To what extent do the substance use patterns and demographic distributions observed in Twitter discourse from 2019 to 2021 correspond with or differ from the baseline trends reported by the National Center for Drug Abuse Statistics (NCDAS) and/or other research?

In order to evaluate the extent to which the substance use patterns and demographics observed in Twitter discourse align with or deviate from baseline trends reported by NCDAS, we conducted a comparative analysis of both dataset. The NCDAS provides comprehensive annual reports on substance use across various demographics, which serve as a benchmark for understanding broader trends. Our analysis focuses on comparing these established trends with the data extracted from Twitter posts spanning from 2019 to 2021. This comparison aims to identify consistencies or discrepancies in substance use trends and demographic patterns between the two sources.

Figure 5. Substance Usership in Twitter Discourse in 2019, 2020 and 2021



Key Findings from Twitter Discourse:

The user distribution across identified substance types as shown in Figure 5, highlighted Cannabinoids, Stimulants, and Opioids as the top three illicit substances in Twitter discourse. Demographically, Male users dominated all substance types in all studied periods as shown in Figure 6. Across the age group, substance use was observed highest in Teenagers <=18 as shown in Figure 7. Usership in Cannabinoids discussion remained the highest among all with the declining trend in both Adults and Teenagers; Teenagers (<=18) showed declining from 2019 to 2021 by 0.29% and 0.07% respectively, and Adults (>18) showed declining by 0.52% and 0.07% respectively. In both Opioids and Stimulants, Adults aged >=40 were observed highly involved among all aged groups in all studied periods. Teenagers (<=18) were observed declining in Opiods from 2019 to 2021. Similarly, the effect of COVID-19 lockdown was evidenced in Alcohol users profoundly (supported by bi-weekly distribution charts in Figure 9) which increased by 80% in just two weeks after global pandemic was declared on March 15, 2020 .

Comparison with NCDAS Trends

Both the reports from NCDAS and our analysis have highlighted Cannabinoids and Stimulants as the top two illicit drugs in the study period . Although the usership in Opioid didn't account as the top substance in Twitter discourse as in NCDAS reports [4], the pattern, however, shows the declining trend in both of the studies, with the decline from 2019 to 2020 of 8.1% in SAMHSA, 2020 [3], and 25% in our study. Our result shows that usership in opioids was mostly prevalent in Adults (age 30 +) compared to Teenagers (<18). This is likely supported by the overdose deaths report [NIDA 2020], where 75% of overdose deaths in adults were from opiods. Likewise, club drugs are highly renowned to be consumed by young people with higher income settings. The same can be seen in our result, where teenagers were observed highly involved compared to other age groups. The similar trend of substance use in terms of gender was observed from both of the studies, Male were actively involved in all substance types except for a few. The exception in Prescription Medication for instance as shown in Figure 8,

showed a higher prevalence in females in both of the studies.

Trend in Alcohol Users in 2020 from other survey based research

The rise in Alcohol users from our analysis during peak pandemic is supported by multiple research [43, 45]. [43] highlighted Alcohol consumption was observed highest as soon as college was closed during Pandemic Lockdown. In the study, they evidenced that alcohol use was high for the users with Mental Health issues and low for those who received social support during the peak time, but didn't correspond as the time passed. The similar result was observed from our theme analysis detailed in question 4, where both social and mentalhealth themes were observed highly associated with alcohol post during the peak pandemic period March 15, 2020 to March 31, 2020. Likewise, [44] also demonstrated that alcohol consumption was high during peak pandemic, which they associated with covid related stress, followed by availability of alcohol and boredom.

Figure 6. Substance Usership by Gender in Twitter Discourse in 2019, 2020 and 2021

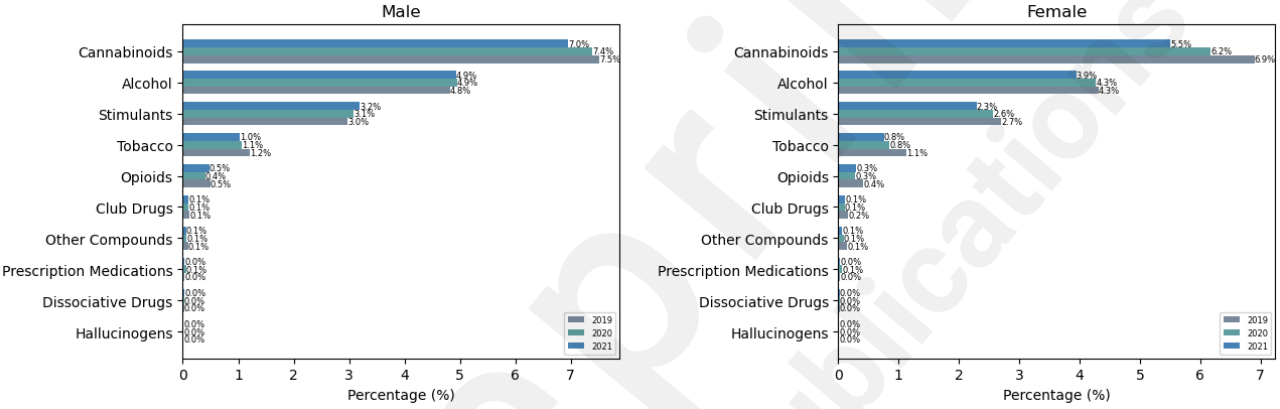


Figure 7. Substance Usership by Age in Twitter Discourse in 2019, 2020 and 2021

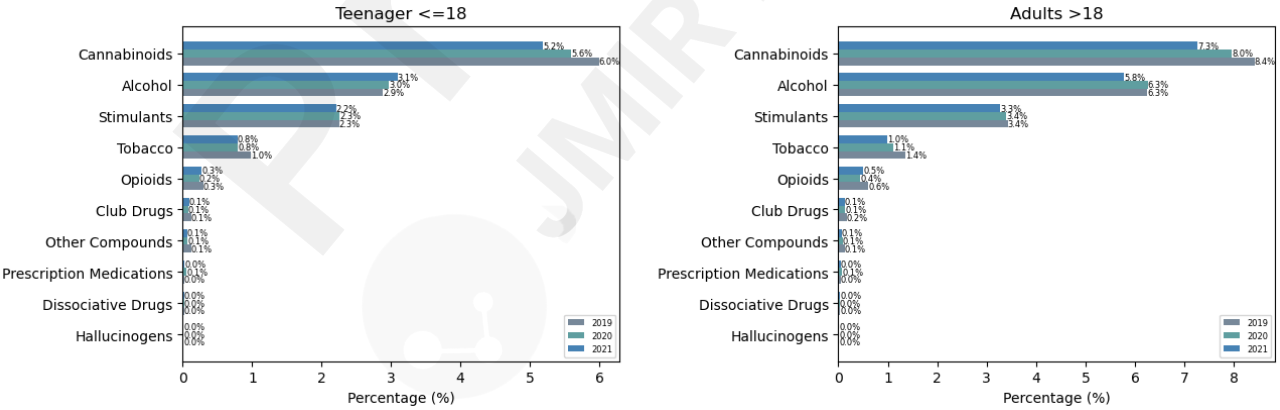


Figure 8. User Demographics in 2020

By Gender

Drug	Female	Male
Cannabinoids	6.17%	7.39%
Alcohol	4.27%	4.95%
Stimulants	2.57%	3.08%
Tobacco	0.85%	1.06%
Opioids	0.28%	0.40%
Club Drugs	0.11%	0.11%
Other Compounds	0.09%	0.07%
Prescription Medications	0.07%	0.06%
Dissociative Drugs	0.02%	0.03%
Hallucinogens	0.01%	0.01%

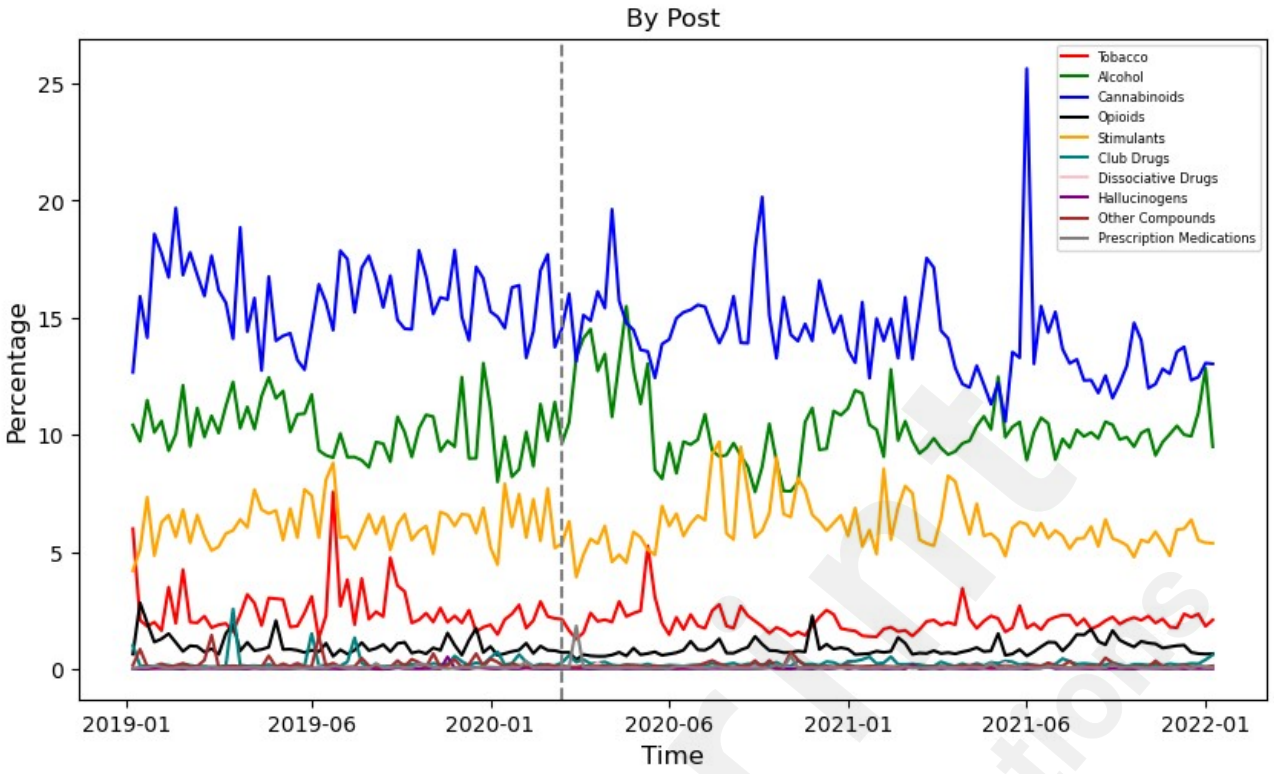
By Age

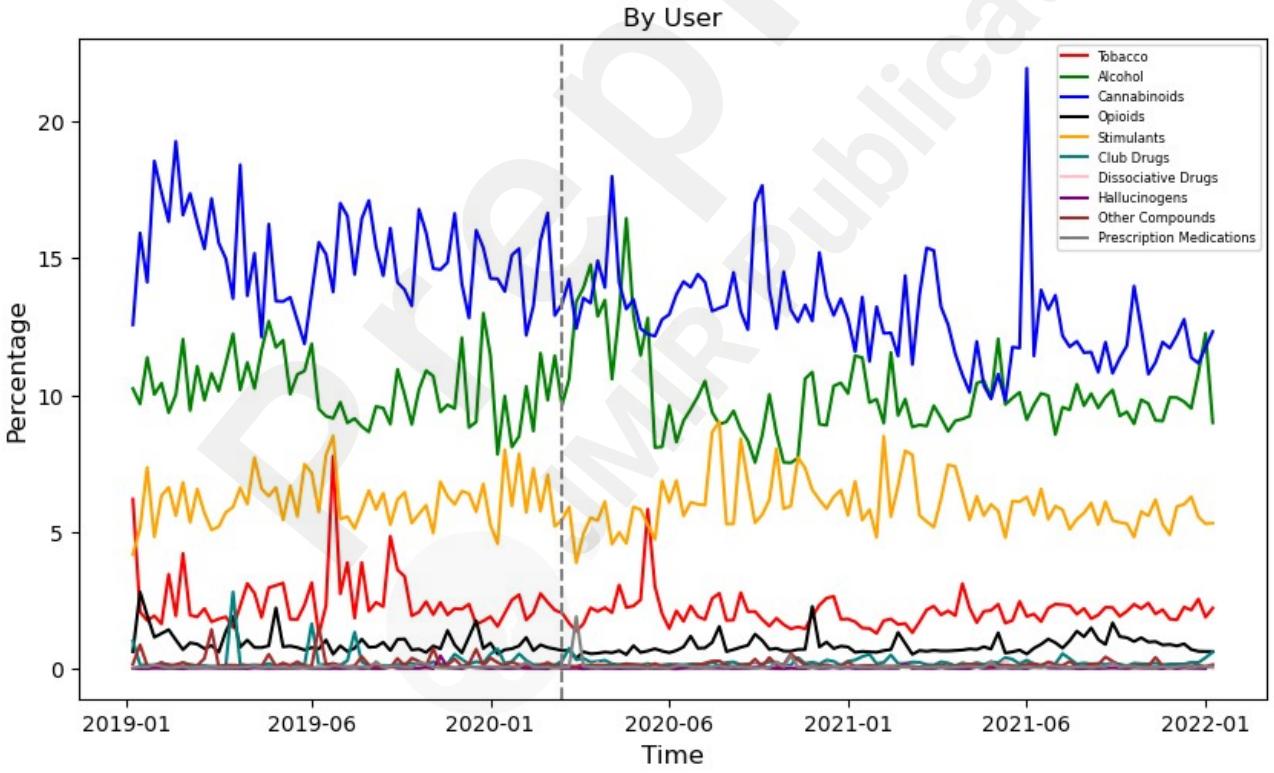
Drug	Teenagers <=18	Adults 18+
Cannabinoids	5.60%	7.96%
Alcohol	2.97%	6.25%
Stimulants	2.26%	3.39%
Tobacco	0.79%	1.11%
Opioids	0.24%	0.44%
Club Drugs	0.07%	0.14%
Other Compounds	0.07%	0.09%
Prescription Medications	0.05%	0.08%
Dissociative Drugs	0.02%	0.03%
Hallucinogens	0.01%	0.01%

Question 3: What are the temporal trends in substance use (SU) posts across different substance types throughout the study period, and how does the frequency of user posting behavior vary over time for each substance type?

The temporal trend analysis gives the nuance of change of proportions with respect to time. In our study, we have plotted weekly trends for all substance types for both users and posts as depicted in Figure 9. At first, we identified substance type for each post using our keyword-based methods, as detailed in our prior work [35]. And to further analyze user posting behavior, we aggregated posts by unique users. Our analysis covers the study period from 2019 to 2021 and presents data on a weekly basis, capturing both short-term fluctuations and long-term trends. The plot highlights Cannabinoids as the most constantly discussed substance among all followed by Alcohol, Stimulants, Tobacco and so forth. While the Cannabinoids posts topped in all study periods, the Alcohol users proportion after Pandemic Declaration day, (Mach 15, 2020, as shown after the dotted gray line in the Figure 7) grew rapidly demanding a detailed focus. Therefore, we present a detailed analysis on alcohol users during this period in our next question.

Figure 9. Weekly Distribution across all Substance Types in Post and User Level





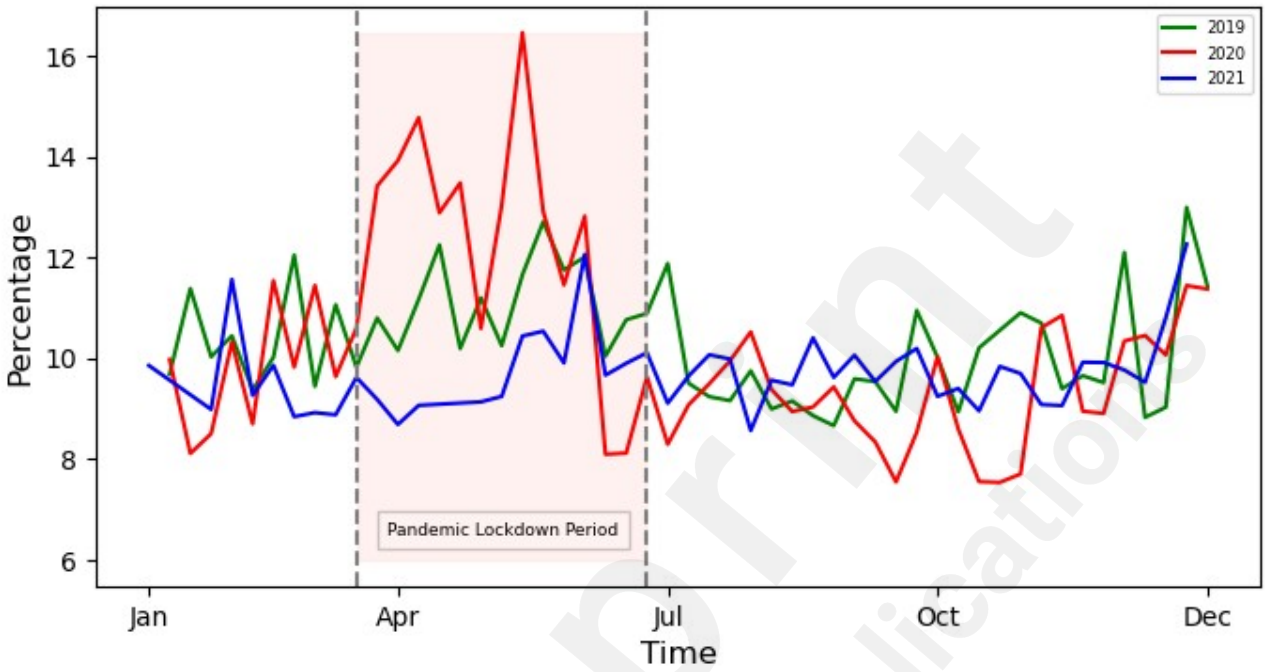
Question 4: How did the number of individuals using Alcohol change in the immediate aftermath of the pandemic declaration compared to those using other substances, and what short-term trends in user behavior across different age groups, genders, races and sentiments during this period?

Trend Analysis on Alcohol Users during Peak Pandemic Period:

Our weekly trend analysis from the questionnaire #3 highlighted that after the pandemic

declaration, Alcohol users surpassed all other substance users including Cannabinoids users (which was the highest discussed substance among all throughout the study period). Hence, we drill down on the Alcohol users to understand if the increase in trend is associated to COVID-19.

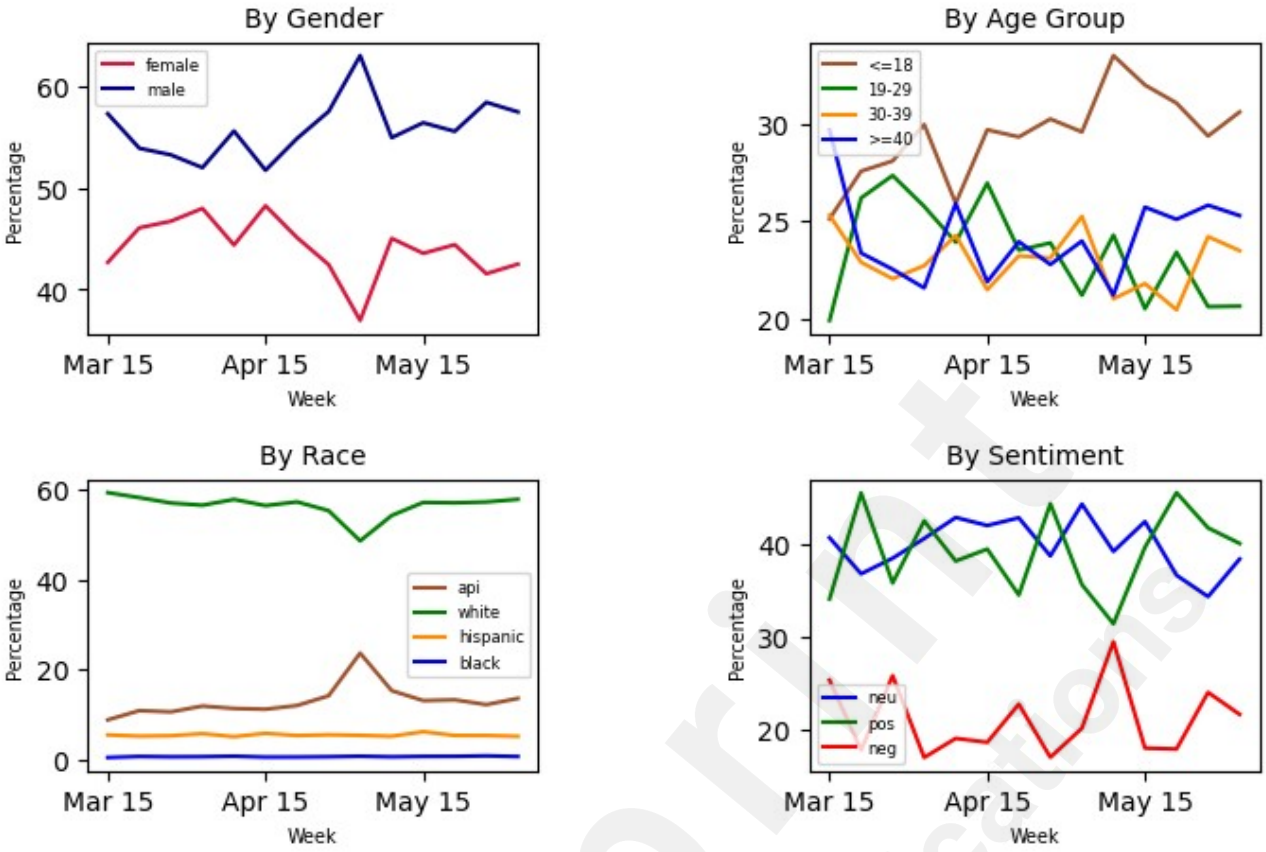
Figure 10. Alcohol: Weekly User Distribution in 2019, 2020 and 2021



Firstly, we compared the users trend (weekly) in the pandemic year with its precursive (2019) and successive (2021) years as shown in Figure 10. The highlighted period from March 15 and June 15 marks the “Pandemic Lockdown Period”. From the visualization, it clarifies that alcohol users dramatically increased during the pandemic year in this period while the proportion in other periods remained intact with its precursive and successive years.

Demographic Trends on Alcohol Users during Peak Pandemic Period:
To further analyze the pattern of alcohol users, we zoomed on the Pandemic Lockdown Period (from March 15, 2020 to Jun15, 2020) in a weekly manner, segmented by age group, gender, race, and sentiment as shown in Figure 11. Each subplot allows for a comparative analysis within these specific demographic or sentiment groups. The gender analysis and Age analysis during this period shows that Male and Teenagers (<=18) users were more involved in Alcohol discussion compared to Female and other age groups respectively. Likewise, for both Male and Teenagers, the trends were seen increasing, and so is for White users from Race Analysis. However, the sentiments during this period were mostly neutral and positive.

Figure 11. User Distribution during Peak Pandemic Lockdown Period (March 15, 2020 to June 15, 2020)



Posts' Content (Theme and Topic) Analysis on Alcohol Use during Peak Pandemic Period:
We performed a detailed analysis on the content of posts, where we derived the underlying themes (Covid, Economic, Social, Mental Health, Supply Disruption, and Medical Disruption) associated with the posts using a keyword method from our previous work [35]. The weekly distribution of alcohol posts in each theme is presented in Figure 12. The distribution showed the alcohol discussion was observed at the highest 2nd week of March (2020-03-15) across all themes except the Economic theme. Further analysis showed that the significant increment in distribution of alcohol discussion was observed in a week of 2020-03-15 specially in Social, MentalHealth, SupplyDisruption and Medical Disruption themes.

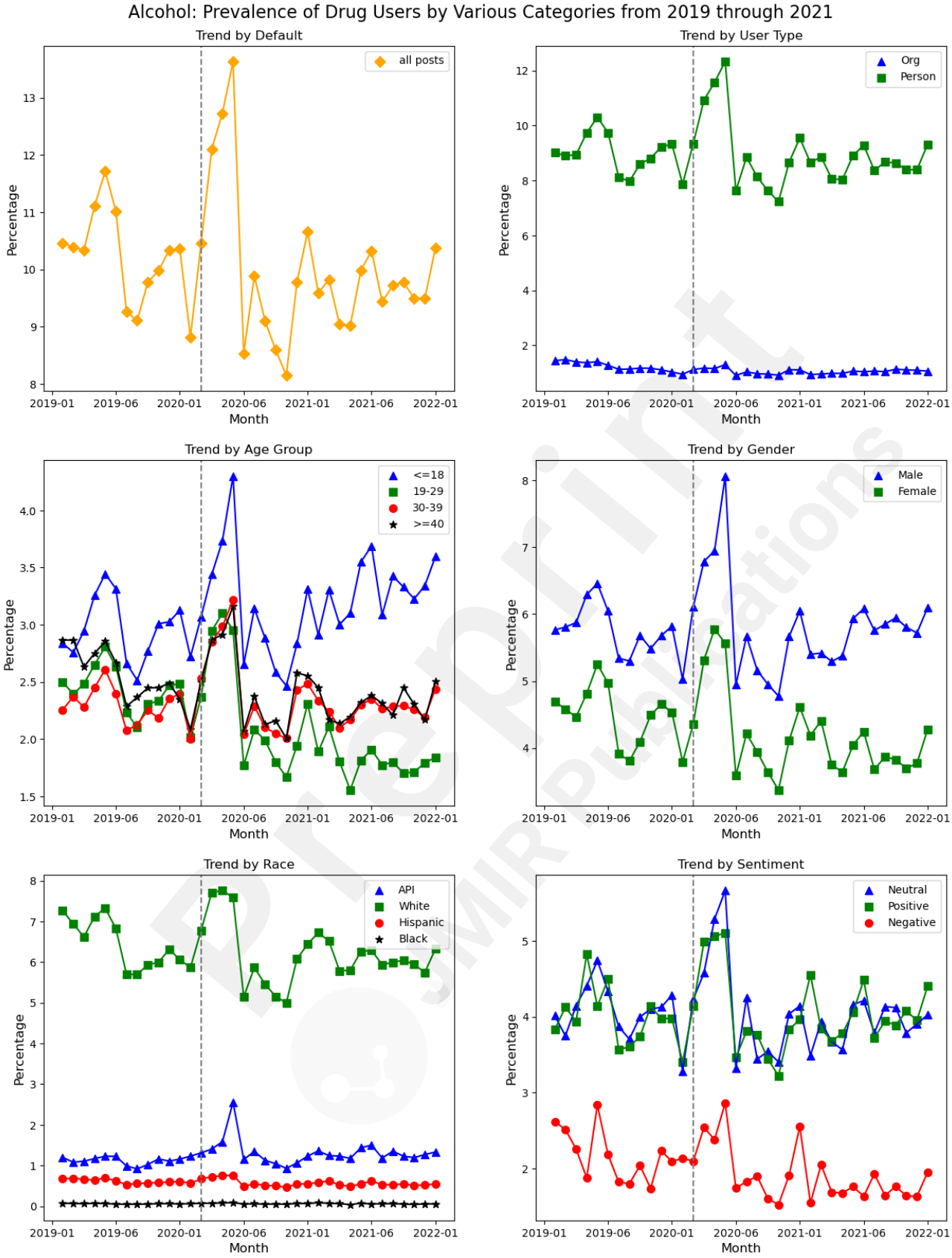
Figure 12. Weekly Alcohol Posts Distribution across all themes from Feb 2020 to May 2020

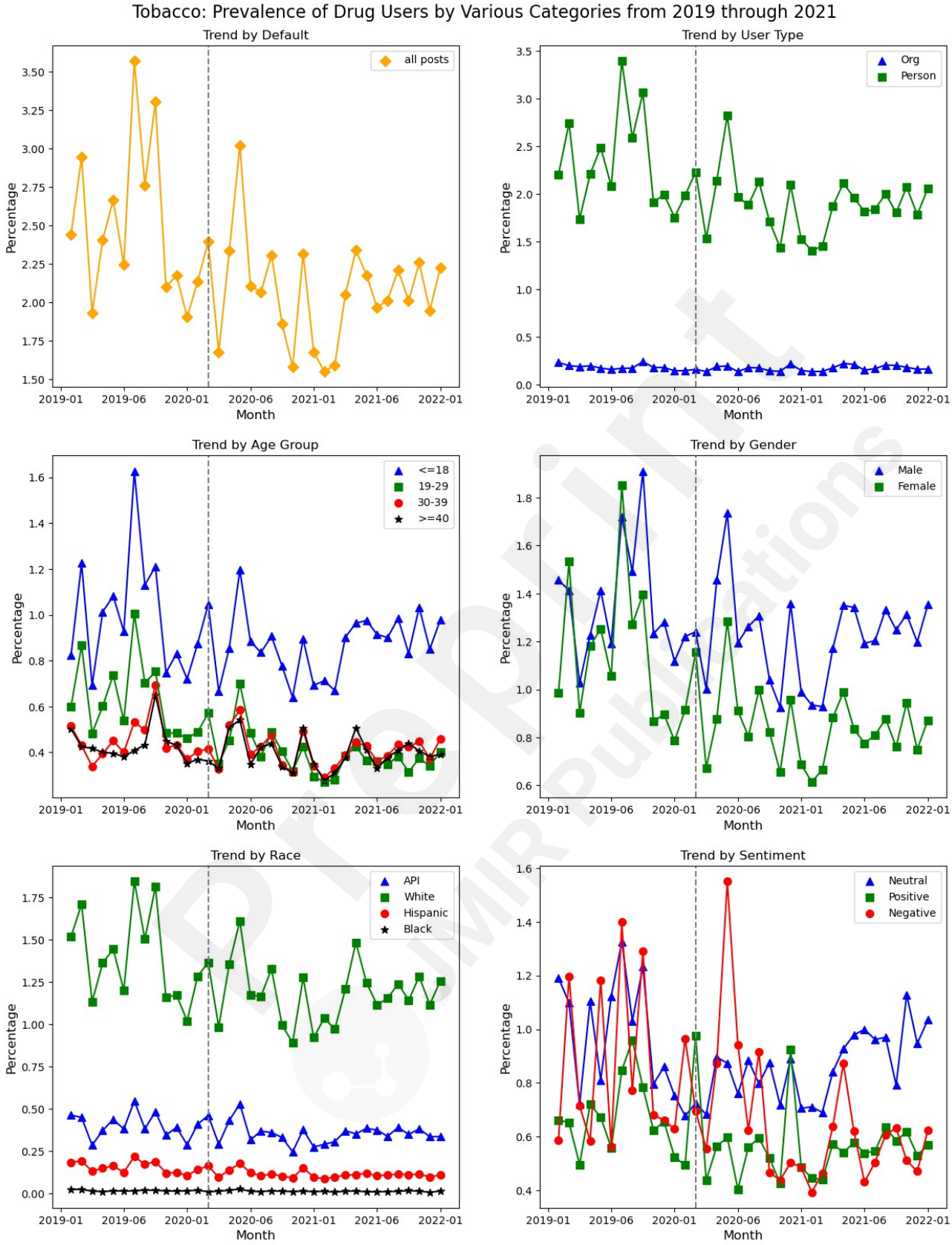
Week	Covid	Economic	Social	Mental Health	Supply Disruption	Medical Disruption
2020-02-01	1.45%	1.81%	0.65%	0.72%	2.85%	0.3%
2020-02-15	1.5%	1.36%	0.71%	0.65%	2.84%	0.32%
2020-03-01	11.15%	4.71%	1.9%	1.22%	2.91%	0.32%
2020-03-15	12.1%	1.61%	8.57%	4.62%	7.4%	4.11%
2020-04-01	5.32%	1.55%	3.4%	1.05%	3.8%	0.63%
2020-04-15	4.86%	1.85%	4.14%	0.94%	3.74%	0.46%
2020-05-01	10.46%	2.88%	3.86%	0.98%	5.57%	0.52%
2020-05-15	2.79%	1.62%	1.58%	0.73%	3.61%	0.45%

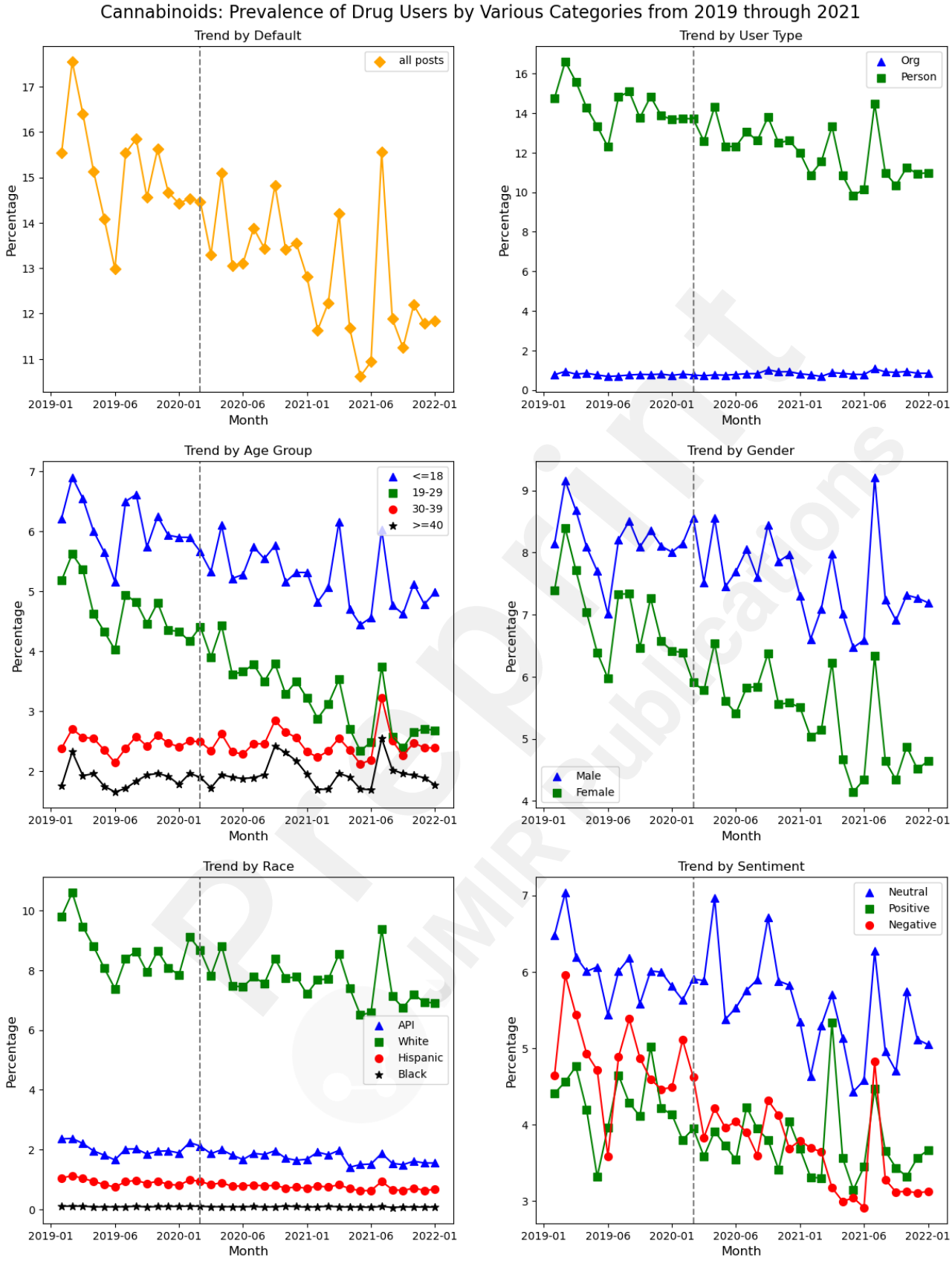
Question 5: What is the trend in Usership across different demographics in each substance type?

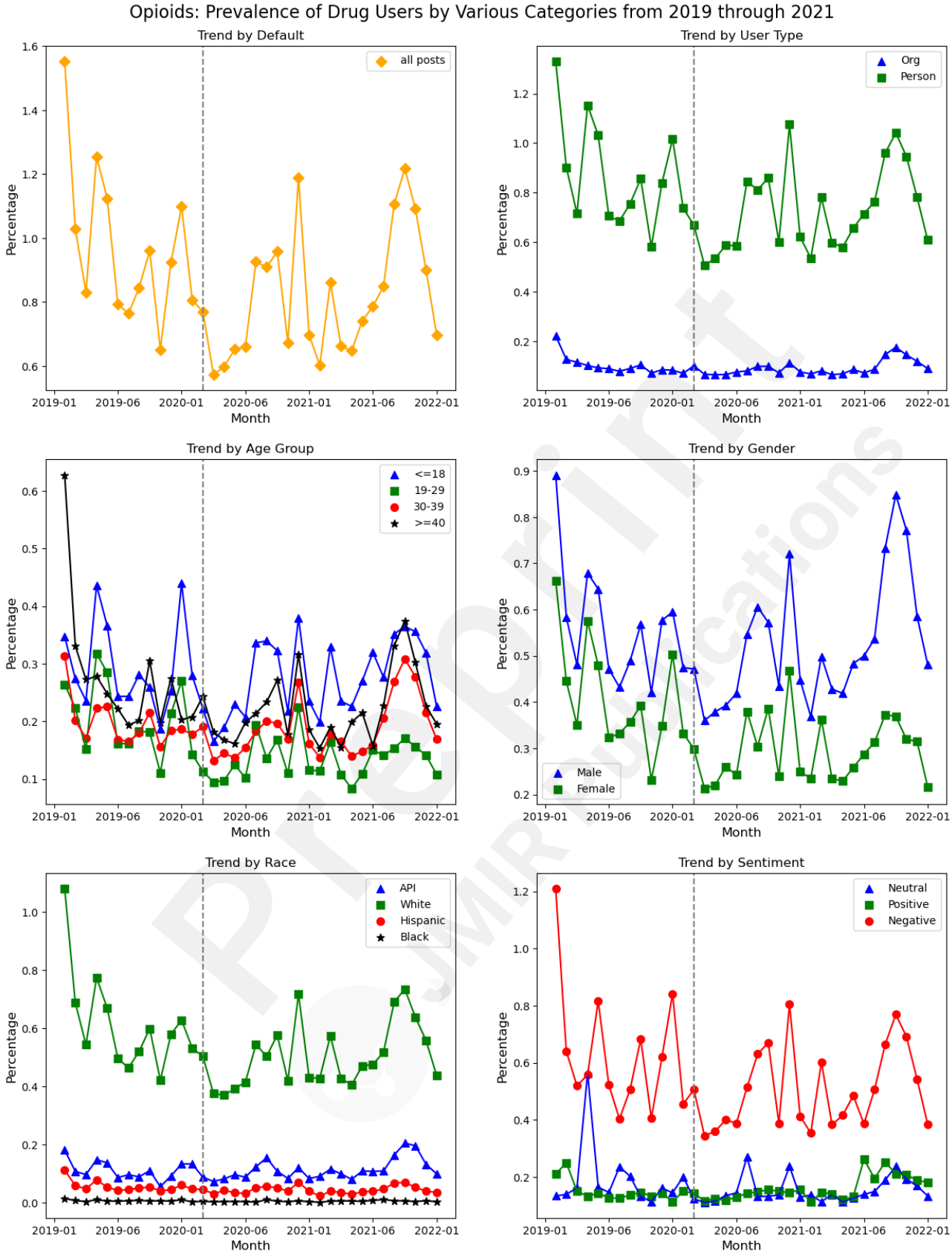
The following diagrams from Figure 13 gives the temporal trend (Monthly) of substance use aggregated by User across 6 dimensions; namely overall (or by default), user type, gender, age, race, and sentiment. The trend aggregated by Post can be found in the Multimedia Appendix Figure 2. The period comprises all 3 years; 2019, 2020 and 2021. The horizontal dotted line in the plots represents the Pandemic Declaration Date to ease visualibility to understand the trend during that period.

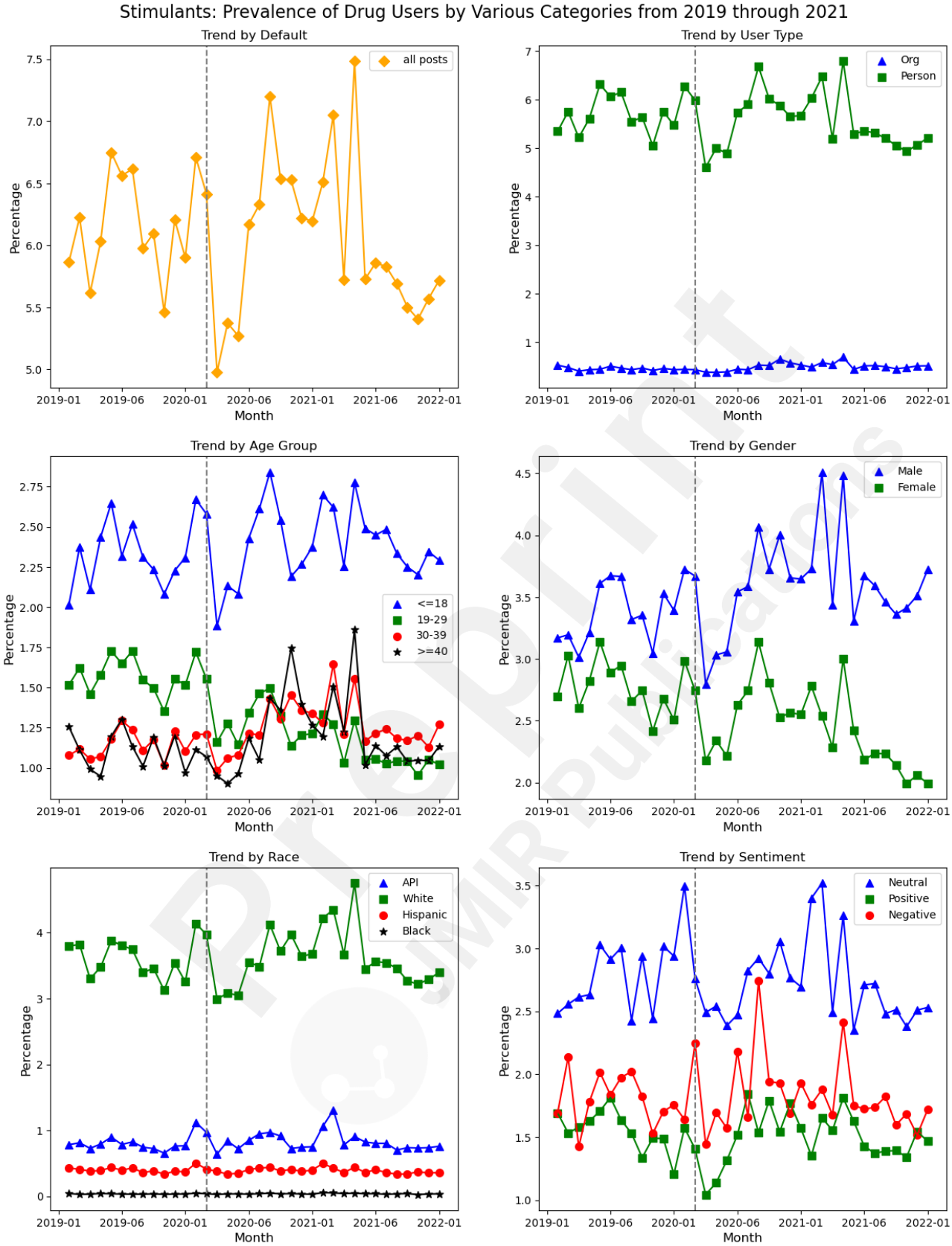
Figure 13. Substance Users distribution across 6 categories from 2019 through 2021

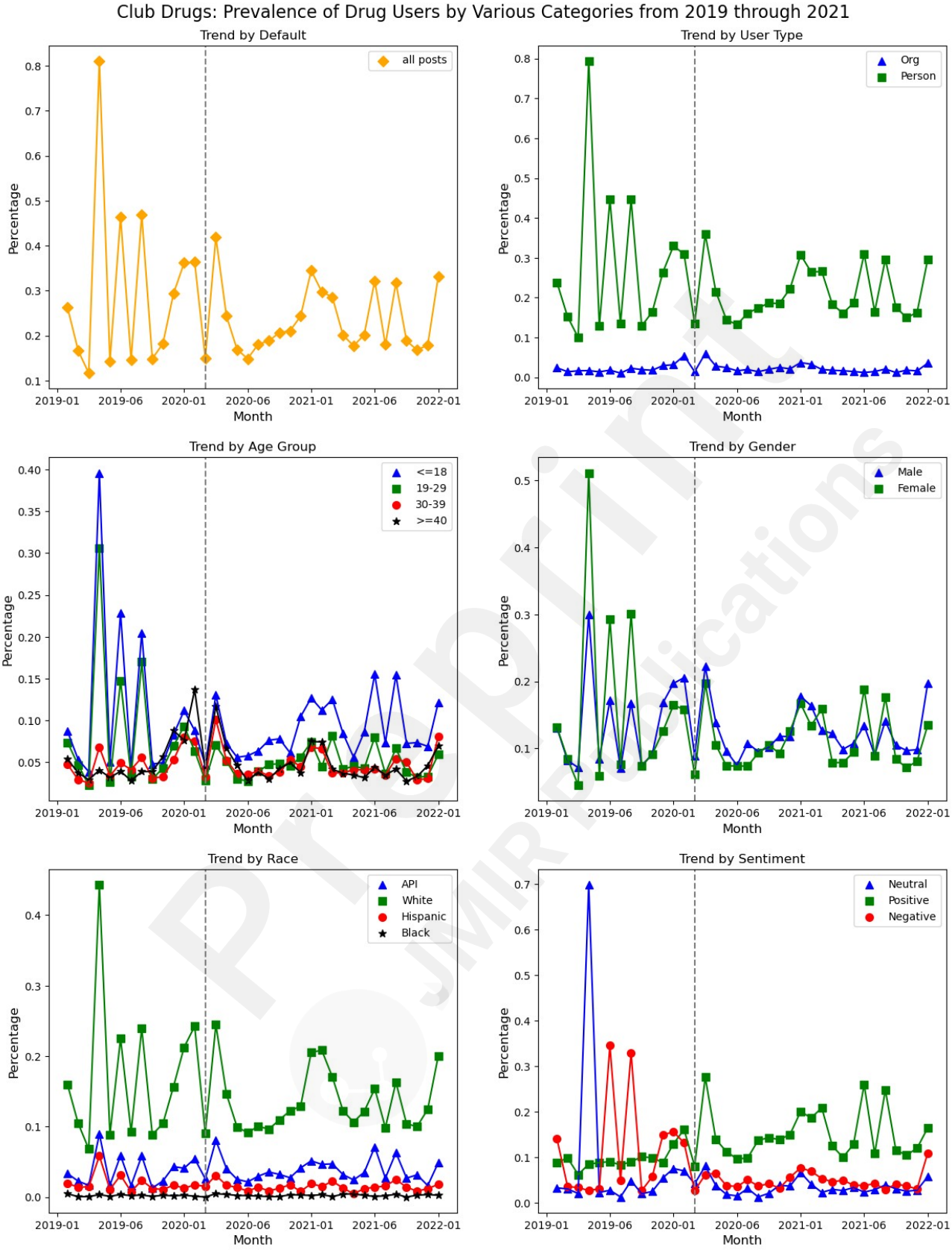

Alcohol: Prevalence of Drug Users by Various Categories from 2019 through 2021
Trend by Default
all posts
Trend by User Type
Org
Person
Trend by Age Group
<=18
19-29
30-39
>=40
Trend by Gender
Male
Female
Trend by Race
API
White
Hispanic
Black
Trend by Sentiment
Neutral
Positive
Negative
Percentage
Month
2019-01
2019-06
2020-01
2020-06
2021-01
2021-06
2022-01

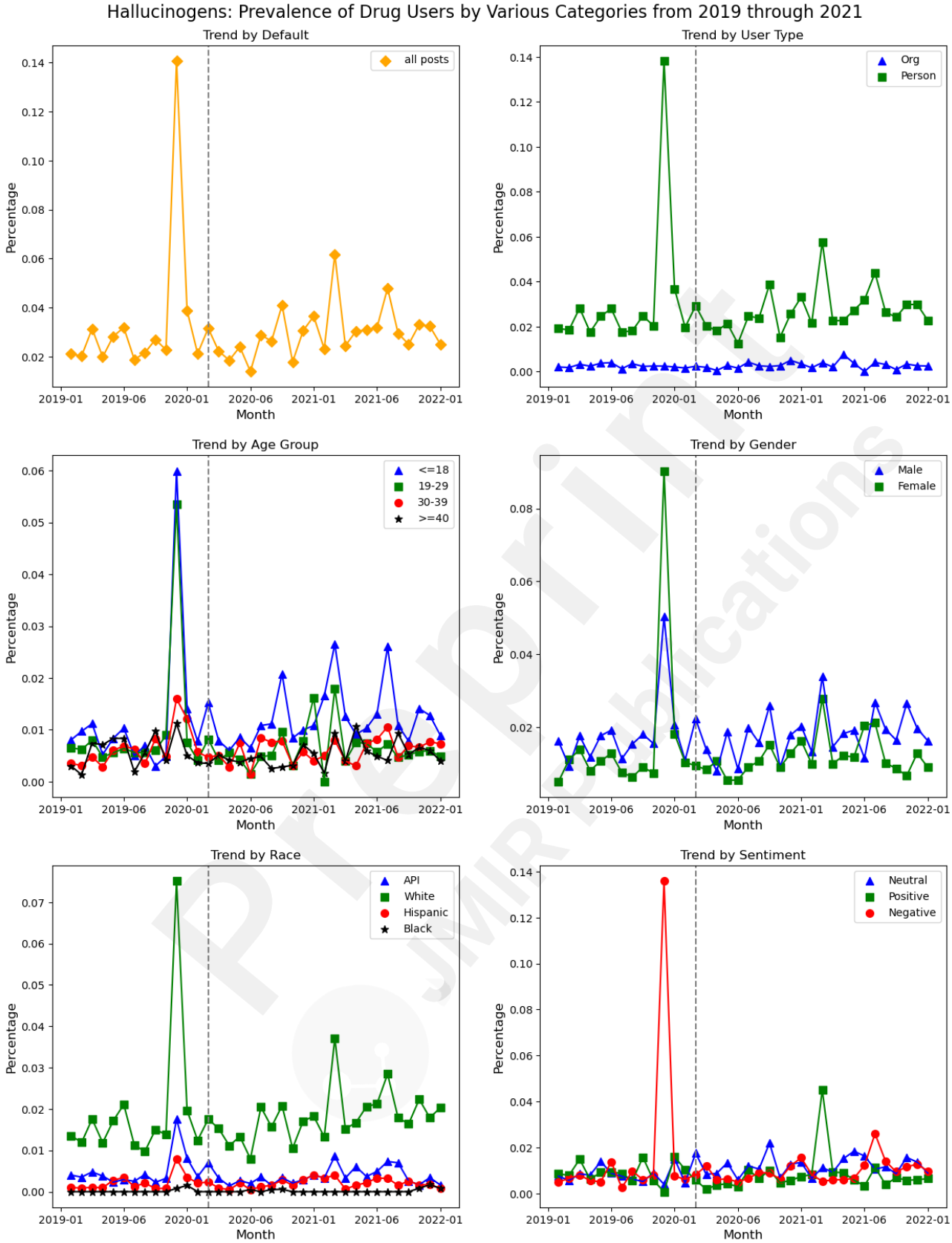


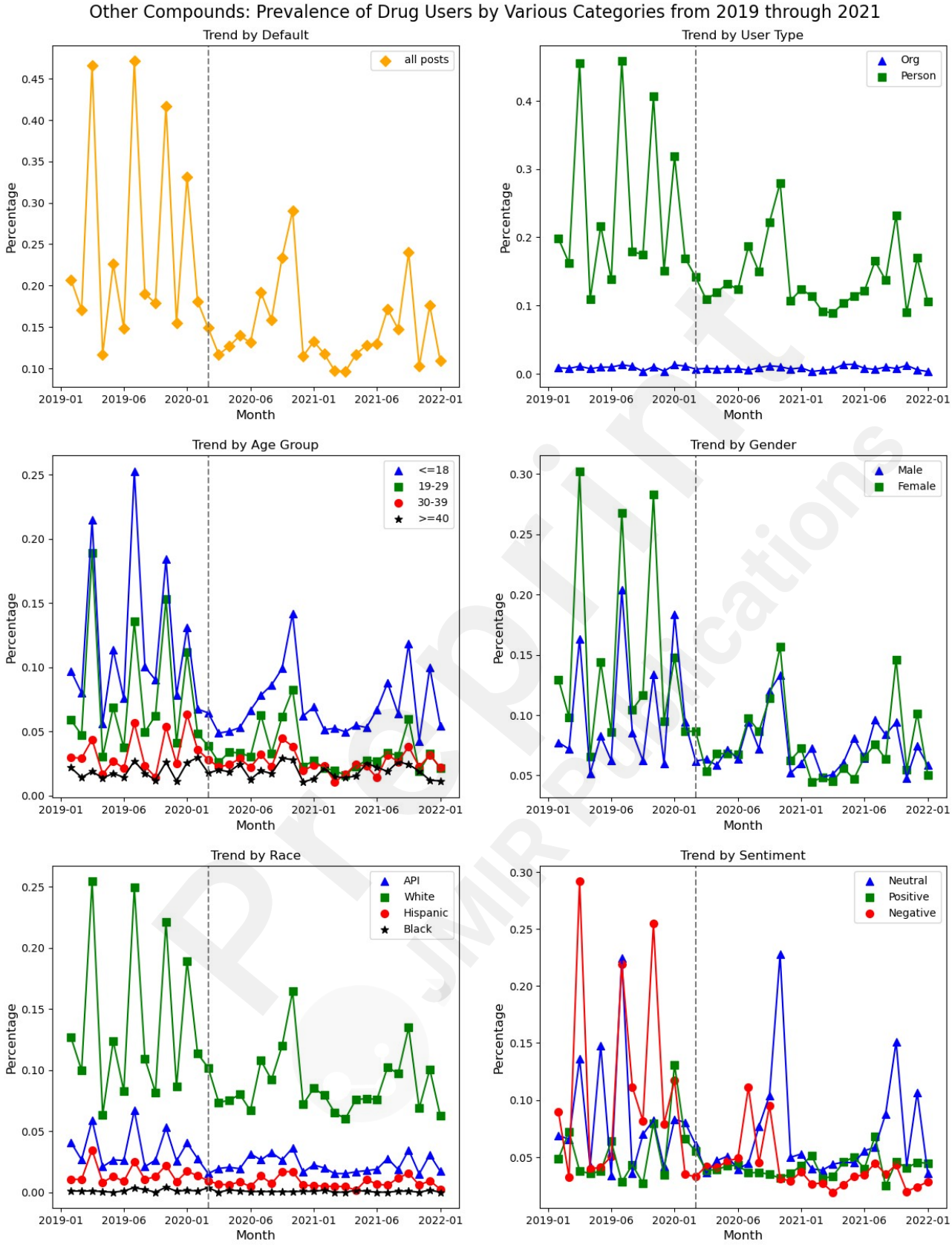


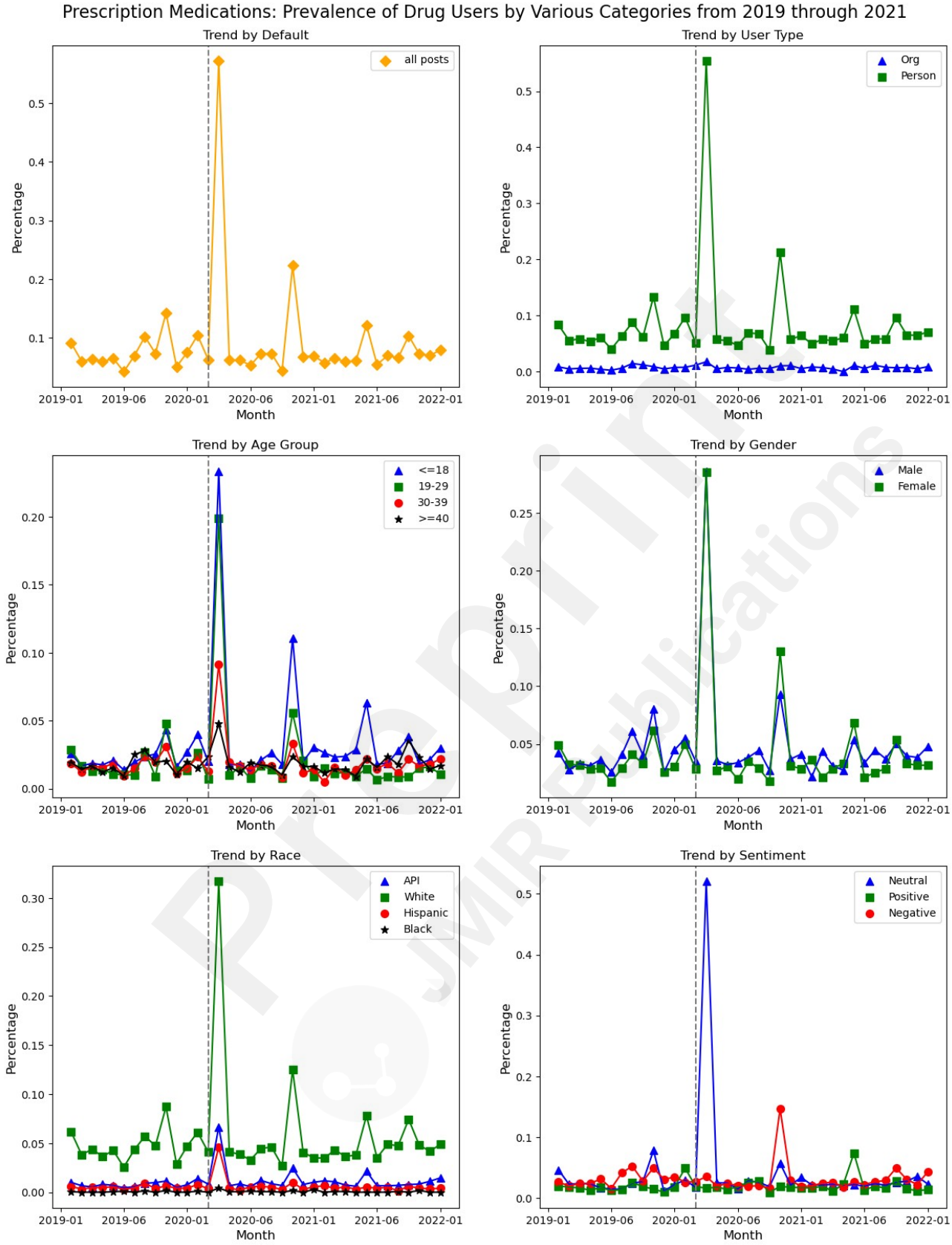








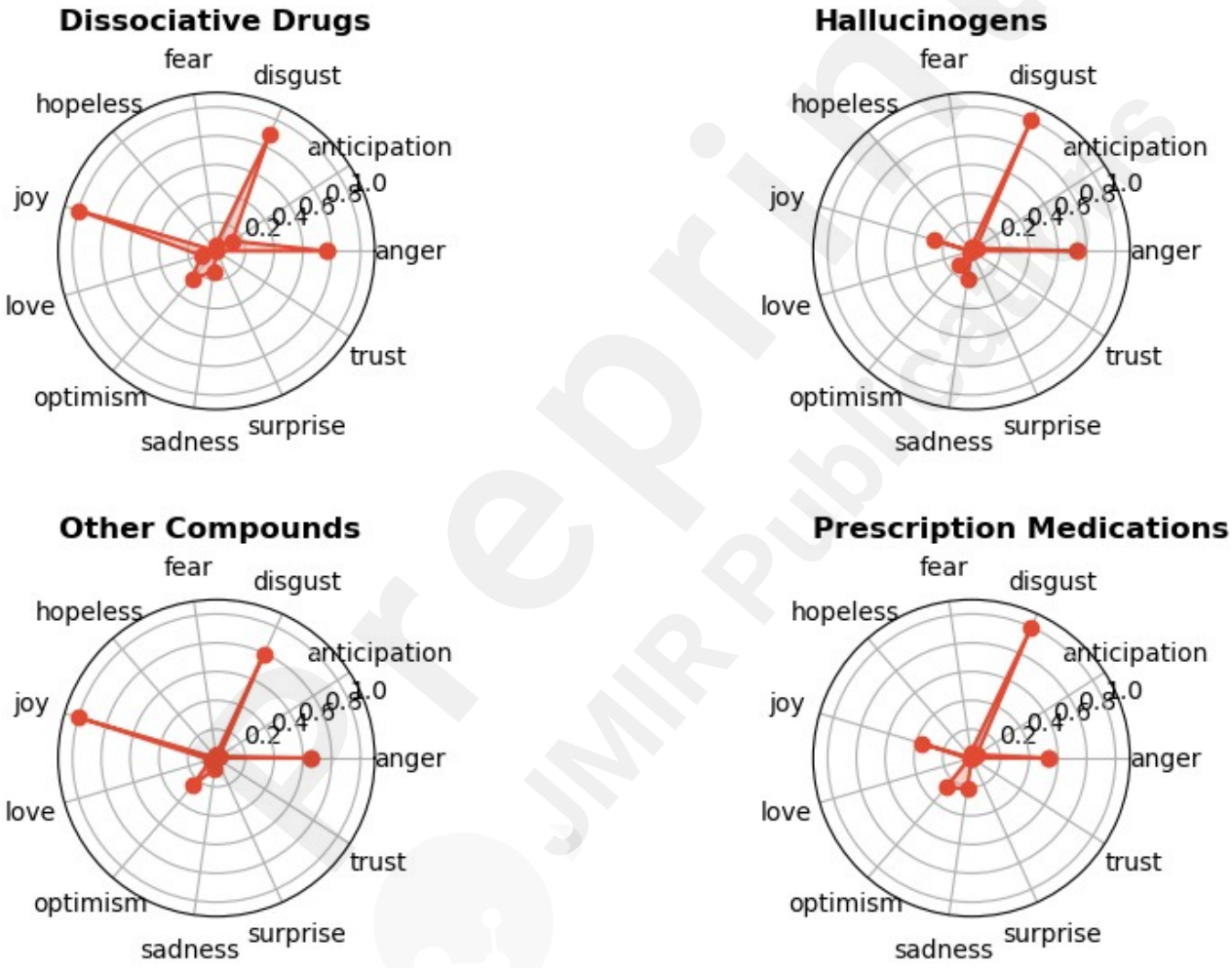


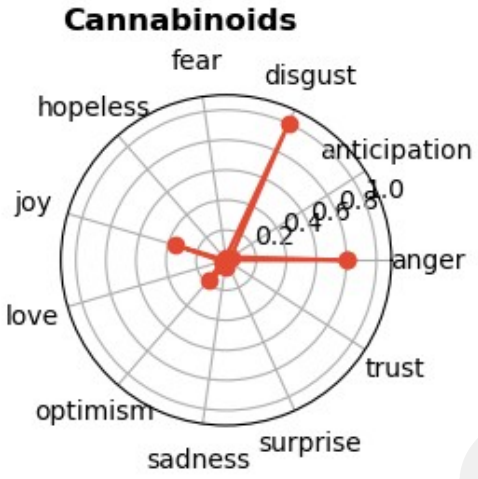


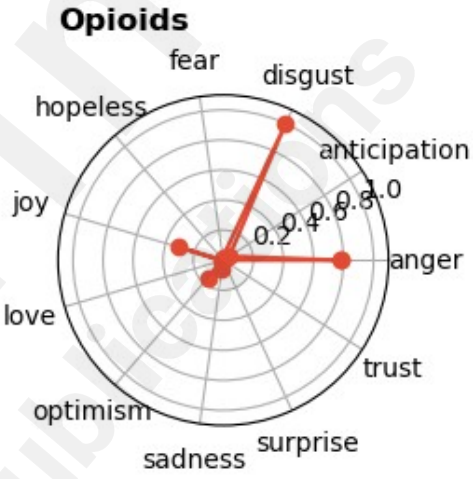
Question 6: What emotional expressions were prevalent in each substance type?

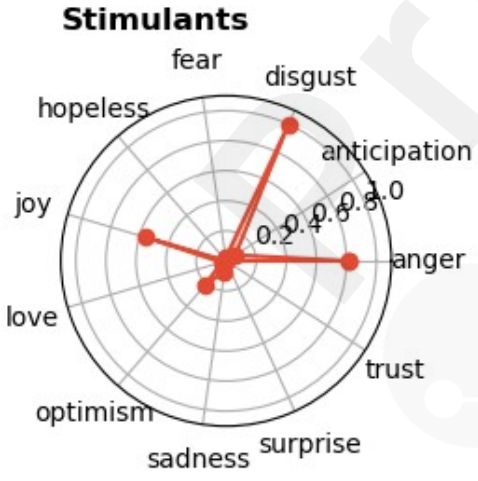
We applied the SpanEmo [38] model to perform emotion detection based on Plutchik Emotion Theory [33] with 10 main emotion categories (i.e, 'anger', 'anticipation', 'disgust', 'fear', 'hopeless', 'joy', 'love', 'optimism', 'sadness', 'surprise', 'trust'). The detected emotions were processed further to calculate the mean intensity scores which are presented in the Radar plot in Figure 14. In our result, we present emotions for each substance type. Our experiment result showed that the majority of the substance discussion mostly incorporates joy, disgust, and anger emotions.

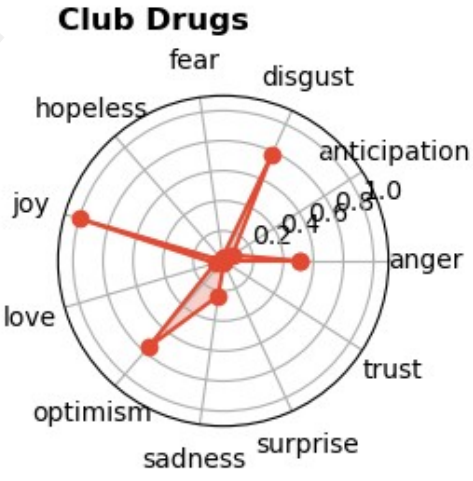
Figure 14. Plutchik Emotion Analysis across all Substance Types











Discussion

Principal Results

Our current study established a foundation for analyzing substance usage across different demographics from an online data perspective, particularly focusing on COVID-19 year. In addition to the substantial findings as a result, we have made a comprehensive comparison

with existing surveys based reports from NCDAS and other research works. Successively, our work has found a notable alignment with survey based reports as discussed in question 2. Notably, users' involvement in substance use-related discussions surged in 2020, with users increasing by 22.18% compared to 2019 and 25.24% compared to 2021. Demographically, male users overtook females in discussions, with their share of posts increasing from 48.87% pre-pandemic to 53.40% post-pandemic, while the youngest age group (<=18) remained the most active, with their proportion growing over time, from 39.56% in 2019 to 41.72% in 2021.
Annually, Cannabinoids, Alcohol, Stimulants, and Tobacco were the top substances (in ascending order) discussed by the online users, while Dissociative Drugs and Hallucinogens were the least. Likewise, the annual trend was observed declining in all top substances, except for Opioids, where 20% decline was observed in only 2020. Similarly, breakdown on age and gender in 2020 revealed that both Adults (>18) and Male users dominated all substances except for few. Prescription Medications and Other Compounds (edible substances) were found to be higher in Female Users, and Tobacco use was observed higher in Teenagers (< =18).
An increase in Alcohol users was observed high following the Global Pandemic Declaration. In just a two week period, the Alcohol users grew by 80%, where the majority of Male Teenagers (<=18) were observed involved in Alcohol discourse, which is also supported by multiple research [43], [44]. Both of the research highlighted Alcohol consumption was observed highest during Peak Pandemic Lockdown period and accounted for mental health, social issues, and covid stress, boredom, availability reasons respectively, which has been evidenced in our theme analysis on alcohol discussion.
Another remarkable pattern was observed in discussion of prescription medication, where females were involved more in social media discourse, which has been also supported by work on the Prescription Drug Misuse and Women by [26], where they found that females are more likely to use prescription medication compared to any other recreational drugs.
Furthermore, our emotion analysis revealed that Alcohol is the only substance that is highly associated with joy emotion, while all other substances are mostly linked with disgust and anger emotions, suggesting positive or happy reasons for consuming alcohol and negative emotions for all other substances.
Overall, our online based research on substance use demonstrates a great potential to bring the insights of global substance use trends under different demographics. Due to the real time nature of unfiltered public opinions, the analysis on social media data can be a great platform to study trends and patterns, specifically during global pandemic when in person collection of data is impossible, to intervene and thus develop prevention strategies in a real time.

Limitations

This study builds upon our initial research. However, missing data from the original source [34] could potentially deviate the actual results. In addition to this, our analysis was limited to English-language posts, which may have excluded non-English speaking users, thus not fully representing the entire user base during the study period. Likewise, the analysis was based on the substance use posts identified using our custom model developed in our previous work [35]. Thus it accounts for only 80% accuracy. Furthermore, demographic extracted were solely based on the profile information like screen name, first name, last name and description of Twitter profile. Thus any variation on this information could mislead the demographic extraction that was carried on. In the extraction of race/ethnicity, there was a tendency to identify larger populations as White, partly due to the focus on English-language posts, as the majority of English language speakers are White. Lastly, the keywords used to identify themes and substance types on the tweet data may have been too narrow, potentially leading to an

overrepresentation of certain themes in our results.

Future work

Our current research has focused on analyzing substance use differences from the perspectives of age, gender, and race. To gain a deeper understanding, future work could incorporate the derivation of post locations to study trends and patterns, thereby allowing for an analysis of the impacting factors specific to different geographic areas. This would enable the development of targeted intervention strategies to prevent substance use based on geographic location. Furthermore, extracting additional information such as socio-economic status and mental/physical health status could significantly enhance the use of social media as a prominent platform for studying public health-related issues. Additionally, analyzing user-based personality traits could provide valuable insights for the public health sector, allowing for the identification of specific characteristics that can inform prevention strategies, even in the absence of demographic information.

Conclusions

Social media platforms collectively with advanced NLP technologies hold a valuable alternative research space that allow researchers to bring insightful aspects of substance use trends and patterns. This study has successfully demonstrated the potential of depicting substance use trends and patterns by aligning the results with the notable survey based reports like NCDAS and MTA. Overall, our result on the Substance Use trend analysis across all demographics provides a baseline study on substance usage that can aid Public Health sectors to focus on specific cohorts to develop efficient interventions thus preventions without need of surveying, during global crises like COVID-19.

Acknowledgments

This work is funded by the Substance Abuse and Mental Health Services Administration SPF-19 Grant (6H79SP081502).

Conflicts of interest

Abbreviations

SU Substance Use
SUD Substance Use Disorder
RoBERTa A Robustly Optimized BERT Pretraining Approach
LDA Latent Dirichlet Allocation
NLP Natural Language Processing
API Asian Pacific Islander
VADER Valence Aware Dictionary and sEntiment Reasoner

References

1. Centers for Disease Control and Prevention. Drug Overdose. URL: https://www.cdc.gov/drugoverdose/featured-topics/recovery-SUD.html [Accessed 2023-08-06]
2. National Institute on Drug Abuse. Covid-19 And Substance Use. URL: https://nida.nih.gov/research-topics/covid-19-substance-use [Accessed 2023-08-14]
3. Substance Abuse and Mental Health Services Administration. Key Substance and Mental Health Indicators in the United States: Results from the 2020 National Survey on Drug Use and Health. URL:

https://www.samhsa.gov/data/sites/default/files/reports/rpt35325/NSDUHFFRPDFWHTMLFiles2020/2020NSDUHFFR1PDFW102121.pdf [Accessed 2023-08-14]
4. National Center of Drug Abuse Statistics. Drug Abuse Statistics. URL: https://drugabusestatistics.org/ [Accessed 2023-10-20]
5. Lee J. Mental health effects of school closures during COVID-19 [published correction appears in Lancet Child Adolesc Health. 2020 Jun;4(6):e16. doi: 10.1016/S2352-4642(20)30128-0]. Lancet Child Adolesc Health. 2020;4(6):421. doi:10.1016/S2352-4642(20)30109-7
6. Addo IY. Double pandemic: racial discrimination amid coronavirus disease 2019. Soc Sci Humanit Open. 2020;2(1):100074. doi: 10.1016/j.ssaho.2020.100074. Epub 2020 Oct 20. PMID: 34173502; PMCID: PMC7574839.
7. Thomeer MB, Moody MD, Yahirun J. Racial and Ethnic Disparities in Mental Health and Mental Health Care During The COVID-19 Pandemic. J Racial Ethn Health Disparities. 2023;10(2):961-976. doi:10.1007/s40615-022-01284-9
8. Chae DH, Yip T, Martz CD, et al. Vicarious Racism and Vigilance During the COVID-19 Pandemic: Mental Health Implications Among Asian and Black Americans. Public Health Rep. 2021;136(4):508-517. doi:10.1177/00333549211018675
9. Gerrard M, Stock ML, Roberts ME, et al. Coping with racial discrimination: the role of substance use. Psychol Addict Behav. 2012;26(3):550-560. doi:10.1037/a0027711
10. Zhu J, Yalamanchi N, Jin R, Kenne DR, Phan N. Investigating COVID-19's Impact on Mental Health: Trend and Thematic Analysis of Reddit Users' Discourse. J Med Internet Res. 2023;25:e46867. Published 2023 Jul 12. doi:10.2196/46867
11. Zhu J, Jin R, Kenne DR, Phan N, Ku WS. User Dynamics and Thematic Exploration in r/Depression During the COVID-19 Pandemic: Insights From Overlapping r/SuicideWatch Users. J Med Internet Res. 2024;26:e53968. Published 2024 May 20. doi:10.2196/53968
12. Cohen S, Wills TA. Stress, social support, and the buffering hypothesis. Psychol Bull. 1985;98(2):310-357.homas Asshby Wills 1990],
13. Czeisler MÉ, Lane RI, Petrosky E, et al. Mental Health, Substance Use, and Suicidal Ideation During the COVID-19 Pandemic - United States, June 24-30, 2020. MMWR Morb Mortal Wkly Rep. 2020;69(32):1049-1057. Published 2020 Aug 14. doi:10.15585/mmwr.mm6932a1
14. Chacon NC, Walia N, Allen A, et al. Substance use during COVID-19 pandemic: impact on the underserved communities. Discoveries (Craiova). 2021;9(4):e141. Published 2021 Dec 31. doi:10.15190/d.2021.20
15. Lee YH, Woods C, Shelley M, Arndt S, Liu CT, Chang YC. Racial and Ethnic Disparities and Prevalence in Prescription Drug Misuse, Illicit Drug Use, and Combination of Both Behaviors in the United States. Int J Ment Health Addict. Published online May 19, 2023. doi:10.1007/s11469-023-01084-0
16. Larochelle MR, Slavova S, Root ED, et al. Disparities in Opioid Overdose Death Trends by Race/Ethnicity, 2018-2019, From the HEALing Communities Study. Am J Public Health. 2021;111(10):1851-1854. doi:10.2105/AJPH.2021.306431
17. Vasilenko SA, Evans-Polce RJ, Lanza ST. Age trends in rates of substance use disorders across ages 18-90: Differences by gender and race/ethnicity. Drug Alcohol Depend. 2017;180:260-264. doi:10.1016/j.drugalcdep.2017.08.027
18. Couch KA, Fairlie RW, Xu H. Early evidence of the impacts of COVID-19 on minority unemployment. J Public Econ. 2020;192:104287. doi:10.1016/j.jpubeco.2020.104287
19. Alcendor DJ. Racial Disparities-Associated COVID-19 Mortality among Minority Populations in the US. J Clin Med. 2020;9(8):2442. Published 2020 Jul 30.

doi:10.3390/jcm9082442
20. Acosta AM, Garg S, Pham H, et al. Racial and Ethnic Disparities in Rates of COVID-19-Associated Hospitalization, Intensive Care Unit Admission, and In-Hospital Death in the United States From March 2020 to February 2021. JAMA Netw Open. 2021;4(10):e2130479. Published 2021 Oct 1. doi:10.1001/jamanetworkopen.2021.30479
21. Johnston, L. D., Miech, R. A., O'Malley, P. M., Bachman, J. G., Schulenberg, J. E., & Patrick, M. E. (2022). Monitoring the future national survey results on drug use, 1975-2021: Overview, key findings on adolescent drug use.
22. Schuler MS, Rice CE, Evans-Polce RJ, Collins RL. Disparities in substance use behaviors and disorders among adult sexual minorities by age, gender, and sexual identity. Drug Alcohol Depend. 2018;189:139-146. doi:10.1016/j.drugalcdep.2018.05.008
23. Daniulaityte R, Carlson R, Falck R, et al. "I just wanted to tell you that loperamide WILL WORK": a web-based study of extra-medical use of loperamide. Drug Alcohol Depend. 2013;130(1-3):241-244. doi:10.1016/j.drugalcdep.2012.11.003
24. Griffiths, M. (2000). Internet addiction-time to be taken seriously?. Addiction research, 8(5), 413-418.
25. Schifano, F., Deluca, P., Baldacchino, A., Peltoniemi, T., Scherbaum, N., Torrens, M., ... & Psychonaut 2002 Research Group. (2006). Drugs on the web; the Psychonaut 2002 EU project. Progress in Neuro-Psychopharmacology and Biological Psychiatry, 30(4), 640-646.
26. Peteet, B., Mosley, C., Miller-Roenigk, B., & McCuistian, C. (2019). Transnational trends in prescription drug misuse among women: A systematic review. International Journal of Drug Policy, 63, 56-73.
27. Murguía, E., Tackett-Gibson, M., & Lessem, A. (Eds.). (2007). Real drugs in a virtual world: Drug discourse and community online. Lexington Books.
28. Griffiths, P., & Mounteney, J. (2010). Drug trend monitoring. Addiction research methods, 337-354.
29. Mounteney, J., Fry, C., McKeganey, N., & Haugland, S. (2010). Challenges of reliability and validity in the identification and monitoring of emerging drug trends. Substance Use & Misuse, 45(1-2), 266-287.
30. Miller, P. G., & Sønderlund, A. L. (2010). Using the internet to research hidden populations of illicit drug users: a review. Addiction, 105(9), 1557-1567.
31. Sarker A, O'Connor K, Ginn R, et al. Social Media Mining for Toxicovigilance: Automatic Monitoring of Prescription Medication Abuse from Twitter. Drug Saf. 2016;39(3):231-240. doi:10.1007/s40264-015-0379-4
32. Correia, R. B., Li, L., & Rocha, L. M. (2016). Monitoring potential drug interactions and reactions via network analysis of instagram user timelines. In Biocomputing 2016: Proceedings of the Pacific Symposium (pp. 492-503).
33. Plutchik, R. (1970). Emotions, evolution, and adaptive processes. In Feelings and emotions: the Loyola Symposium(pp. 1-14). New York: Academic Press.
34. Twitter. 2019 to 2021: Twitter: free download, borrow, and streaming. Internet Archive. [Accessed Oct 2022]
35. Maharjan J, Zhu J, King J, Phan H, Kenne D, Jin R Large-Scale Analysis of Substance Use Trends during COVID-19: An Advanced Deep Learning Approach J Med Internet Res 2024;0:e0 URL: https://preprints.jmir.org/preprint/59076 doi: 10.2196/59076 [Under Review]
36. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Published online 2017. doi:10.48550/ARXIV.1706.03762

37. Yinhan Liu and Myle Ott and Naman Goyal and Jingfei Du and Mandar Joshi and Danqi Chen and Omer Levy and Mike Lewis and Luke Zettlemoyer and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv 2019 Jul 2019 doi: https://doi.org/10.48550/arXiv.1907.11692
38. Wang, Z., Hale, S., Adelani, D. I., Grabowicz, P., Hartman, T., Flöck, F., & Jurgens, D. (2019, May). Demographic inference and representative population estimates from multilingual social media data. In The world wide web conference (pp. 2056-2067).
39. Hutto, C., & Gilbert, E. (2014, May). 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 (Vol. 8, No. 1, pp. 216-225).
40. Alhuzali, H., & Ananiadou, S. (2021). SpanEmo: Casting multi-label emotion classification as span-prediction. arXiv preprint arXiv:2101.10038.
41. Parasurama, P. (2021). raceBERT--A Transformer-based Model for Predicting Race and Ethnicity from Names. arXiv preprint arXiv:2112.03807.
42. Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: a state of the art. Artificial Intelligence Review, 56(4), 3005-3054.
43. Census Data. 2010 Census Data Products: United States. URL: https://www.census.gov/ [Accessed 2024-04-27]
44. Lechner, W. V., Laurene, K. R., Patel, S., Anderson, M., Grega, C., & Kenne, D. R. (2020). Changes in alcohol use as a function of psychological distress and social support following COVID-19 related University closings. Addictive behaviors, 110, 106527.
45. Grossman, E. R., Benjamin-Neelon, S. E., & Sonnenschein, S. (2022). Alcohol consumption and alcohol home delivery laws during the COVID-19 pandemic. Substance abuse, 43(1), 1141-1146.
46. Stacy Jo Dixon (2023). X/Twitter: number of users worldwide 2024 | Statista Research Department [Accessed 2024-08-28]
47. JulinaM/DemographicAnalysis. Github. URL: https://github.com/JulinaM/DemographicAnalysis

Multimedia Appendix 1

Table 1. Gender, Age Group, and Org prediction result for sample posts (with Name, Description, Screen name) from M3Inference

Name	Screen name	Description	Age Group	Gender	Org
Elon Musk	elonmusk	Nan	>=40	Male	non-org
Barak Obama	BarackObama	Dad, husband, President, citizen.	>=40	Male	non-org
Mckenna Grace	MckennaGracefull	Instagram:@mckennagraceful	<=18	Female	non-org
Millie Bobby Brown	Milliestopshate	I want this account to share love and positivity	<=18	Female	non-org
NASA	NASA	There's space for	NAN	NAN	org

		everybody			

Table 2. Race prediction result for sample posts (First name and Last name)from Ethicolr

Name	**Race**
Michael Jackson	Black
John Smith	White
Austin Alderson	White
Zhang	API
Shakya	API
Kendrick Valdez	Hispanic

Table 3. Sentiment prediction result for sample posts from VADER

Tweet	**Sentiment**
USER stag drunkenness buzz tipsy drunk hammer nancy pelosi	Negative
USER night jump drunk men try brake fight break cartilage nose bone	Negative
USER yeah smoker notice smell	Positive
USER yeah smoker notice smell	Positive
USER thinking nicotine	Neutral
USER girl texas tech lose life week ago someone thought could drive drunk mom amp dad lose daughter	Neutral

Table 4. Top 5 words associated with each emotion class

Negative emotions	
Anger	death, think, public, virus, don’t, against
Disgust	deaths, virus, against, because, public, after
Fear	deaths, spread, symptoms, coronavirus, identify, self-reporting
Sadness	deaths, going, cases, hospital, other, please
Pessimism	sadly, family, friend, during, weeks, passed
Positive emotions	

Anticipation	support, vaccine, first, working, public, cases
Joy	great, thank, support, happy, amazing, staysafe
Trust	trust, thank, protect, important, community, everyone
Love	happy, loved, share, beautiful, wonderful, amazing
Optimism	please, thank, support, working, great, spread
Surprise	shocking, surprised, amazing, public, absolutely, deaths

Table 5. Sample prediction result for posts from SpanEmo

Tweet	**Emotions**
@Adnan__786__ @AsYouNotWish Dont worry Indian army is on its ways to dispatch all Terrorists to Hell	['anger', 'disgust', 'fear']
Academy of Sciences, eschews the normally sober tone of scientific papers and calls the massive loss of wildlife a “biological annihilation	['anger', 'disgust', 'sadness']
I blew that opportunity -__- #mad	['anger', 'disgust']

Table 6. Distribution of Substance Use Discourse Aggregated by Post and User

	Pre-Pandemic (2019)		**During Pandemic (2020)**		**Post-Pandemic (2021)**	
	By Posts	**By Users**	**By Posts**	**By Users**	**By Posts**	**By Users**
Total SU Counts	2,799,726	2,131,457	3,502,171	2,604,123	2,553,235	1,946,742
User type						
Org	221,934 (7.93%)	153,779 (7.21%)	291,136 (8.31%)	195,648 (7.51%)	232,207 (9.09%)	159,816 (8.21%)
Person	2,577,792 (92.07%)	1,977,678 (92.79%)	3,211,035 (91.69%)	2,408,475 (92.49%)	2,321,028 (90.91%)	1,786,926 (91.79%)
Gender Type						
Female	1,318,063 (51.13%)	1,007,368 (50.94%)	1,558,219 (48.53%)	1,166,535 (48.43%)	1,081,682 (46.60%)	828,102 (46.34%)
Male	1,259,729 (48.87%)	970,310 (49.06%)	1,652,816 (51.47%)	1,241,940 (51.57%)	1,239,346 (53.40%)	958,824 (53.66%)
Age Group						
<=18	1,019,817 (39.56%)	793,067 (40.10%)	1,257,269 (39.15%)	970,641 (40.30%)	968,414 (41.72%)	759,859 (42.52%)
19-29	772,388 (29.96%)	587,274 (29.70%)	849,019 (26.44%)	633,760 (26.31%)	542,770 (23.38%)	419,018 (23.45%)

30-39	468,354 (18.17%)	361,423 (18.28%)	638,334 (19.88%)	476,802 (19.80%)	488,024 (21.03%)	370,951 (20.76%)
>=40	317,233 (12.31%)	235,914 (11.93%)	466,413 (14.53%)	327,272 (13.59%)	321,820 (13.87%)	237,098 (13.27%)
Sentiment						
Neutral	985,973 (35.22%)	752,499 (35.30%)	1,254,059 (35.81%)	933,994 (35.87%)	938,201 (36.75%)	720,580 (37.01%)
Positive	814,611 (29.10%)	622,641 (29.21%)	1,006,366 (28.74%)	748,546 (28.74%)	788,800 (30.89%)	592,765 (30.45%)
Negative	999,142 (35.69%)	756,317 (35.48%)	1,241,746 (35.46%)	921,583 (35.39%)	826,234 (32.36%)	633,397 (32.54%)
Race*						
Total Race identified	1,811,516 (64.70%)	1,811,516 (64.70%)	2,275,943 (64.98%)	2,275,943 (64.98%)	1,723,470 (67.50%)	1,723,470 (67.50%)
API	303,500 (11.77%)	272,093 (13.76%)	406,301 (12.65%)	363,338 (15.09%)	302,074 (13.01%)	277,011 (15.50%)
White	1,215,550 (47.15%)	1,114,892 (56.37%)	1,506,359 (46.91%)	1,367,638 (56.78%)	1,142,020 (49.20%)	1,053,655 (58.96%)
Hispanic	126,121 (4.89%)	117,406 (5.94%)	148,930 (4.64%)	136,620 (5.67%)	107,018 (4.61%)	99,447 (5.57%)
Black	12,596 (0.49%)	11,731 (0.59%)	16,045 (0.50%)	14,879 (0.62%)	12,068 (0.52%)	11,299 (0.63%)
Unidentified	988,210 (35.29%)	988,210 (35.29%)	1,226,228 (35.01%)	1,226,228 (35.01%)	829,765 (32.49%)	829,765 (32.49%)

Note:

1. The **Total SU Counts** identified varied across years due to fluctuations in the number of posts and users. Consequently, changes in proportions do not necessarily reflect a consistent change in the weighted count.
2. The notation (*) in Race* indicates that the proportion within the subgroup does not account for the base total of tweets. This is because not all posts included a valid username that could be used to infer race or ethnicity, leading to a reduction in the total number of posts after race identification.

Figure 1: Alcohol Distribution by demographics in 2019, 2020 and 2021

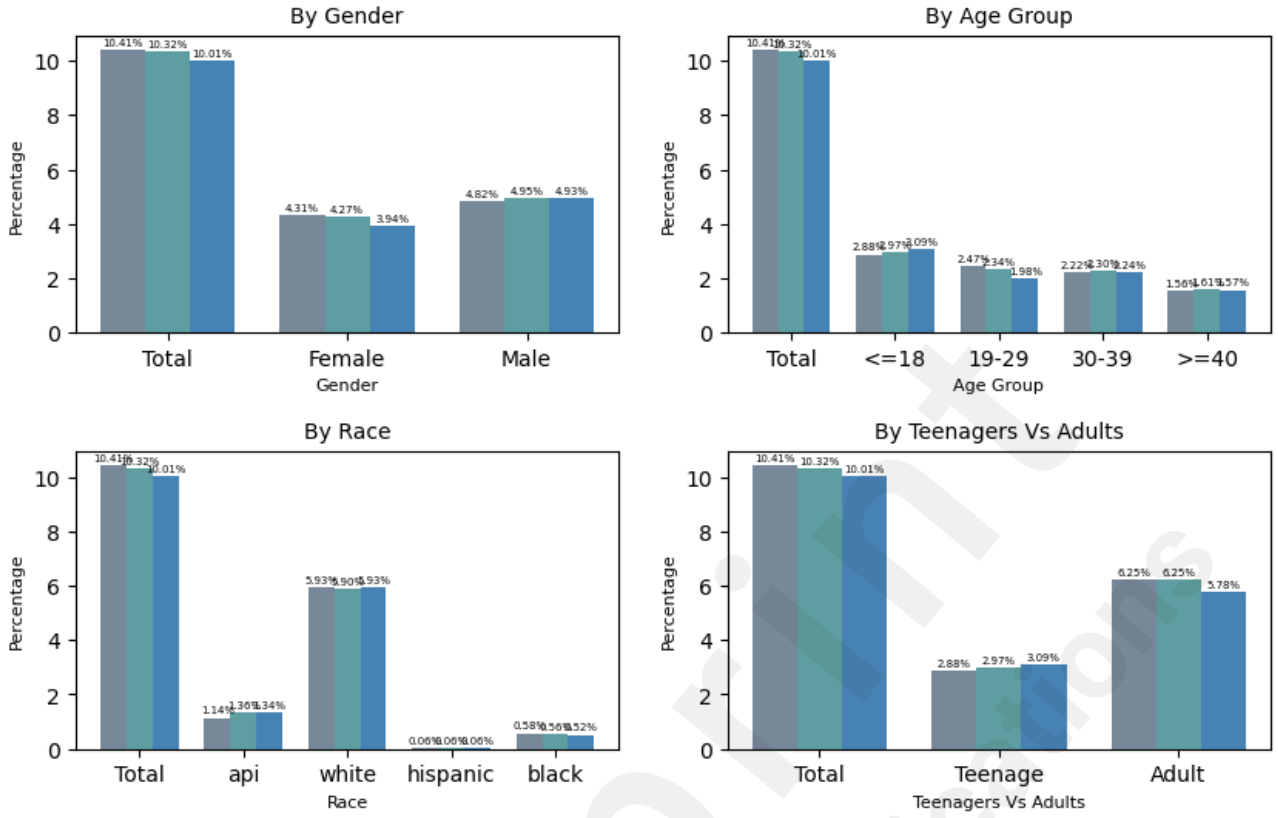
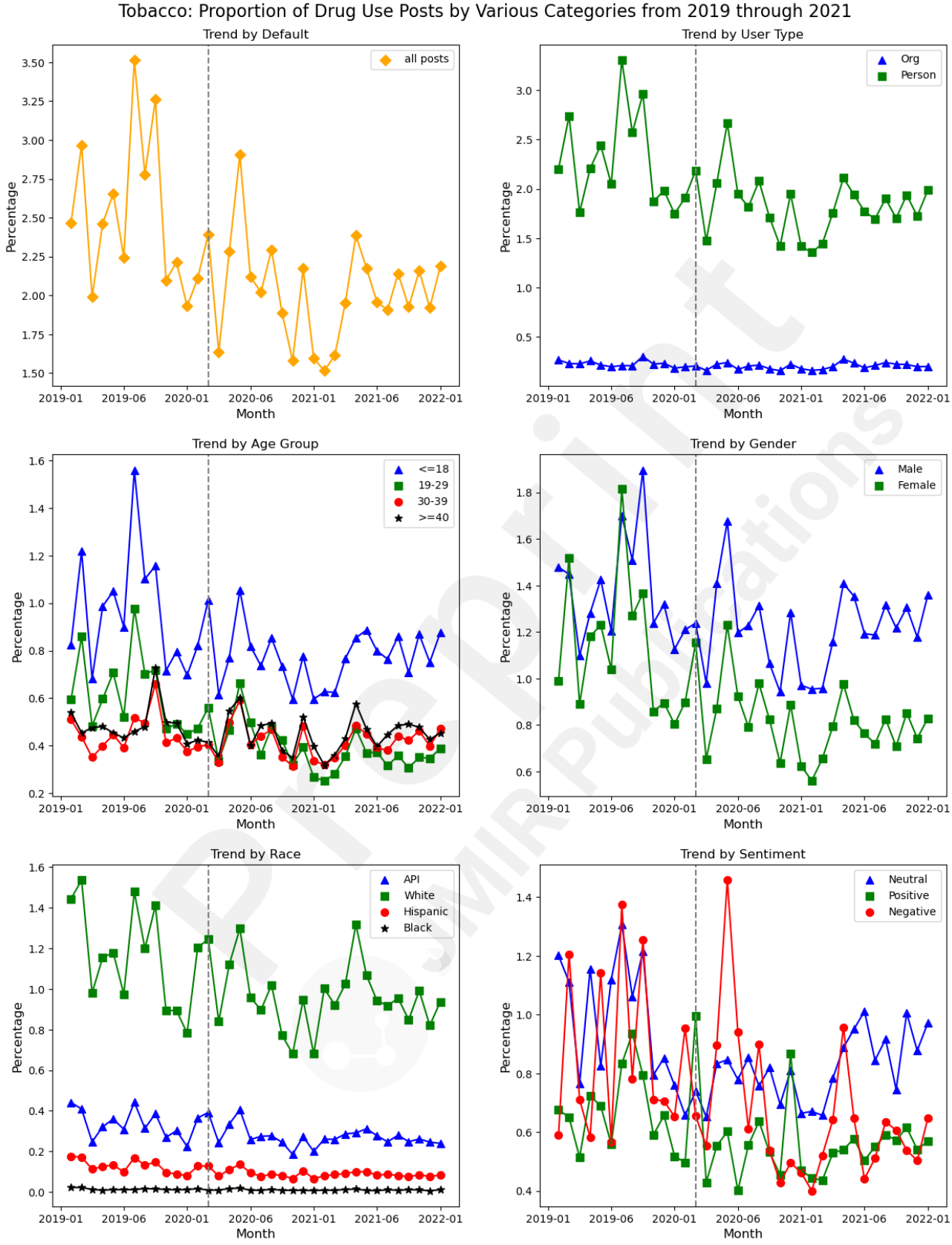
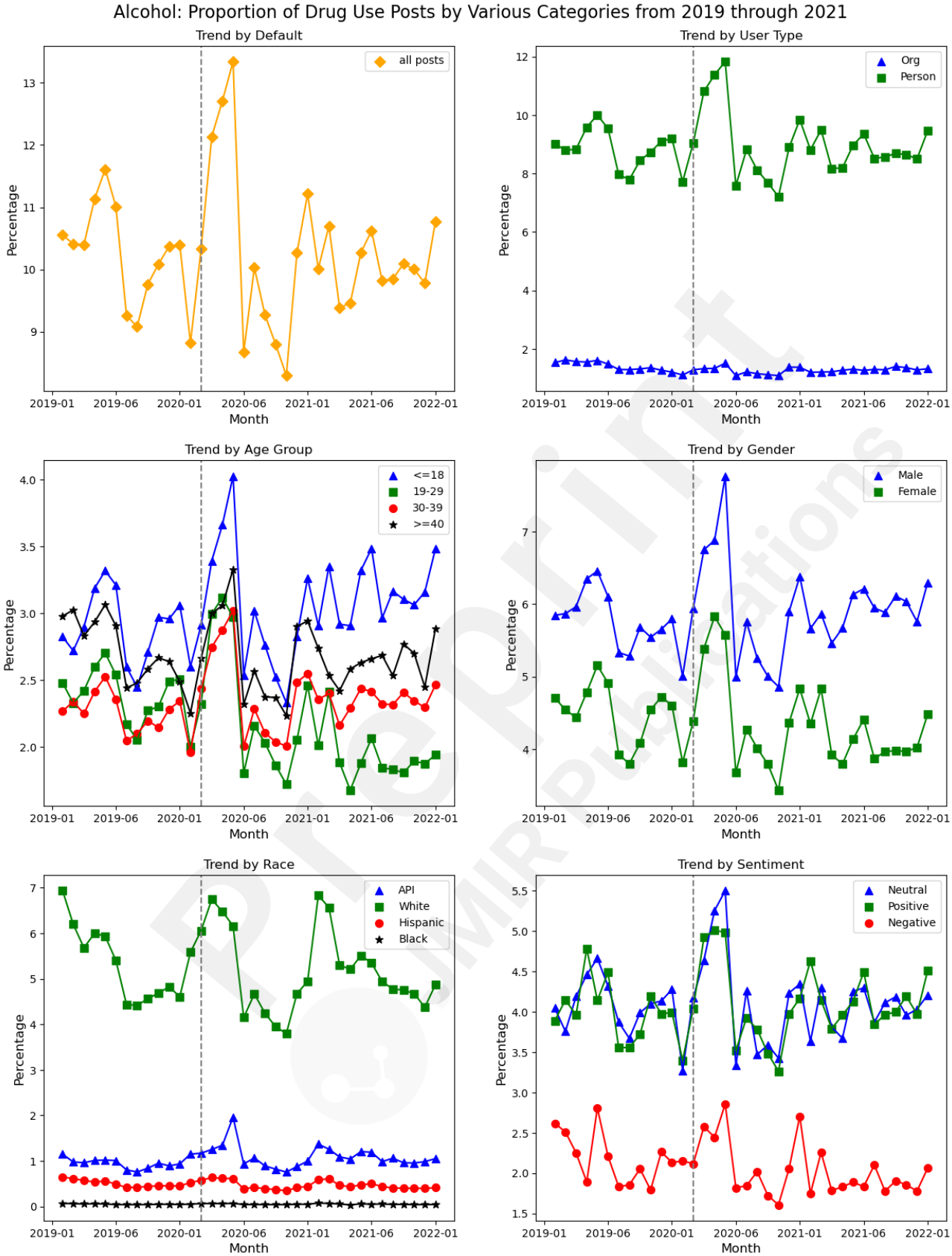
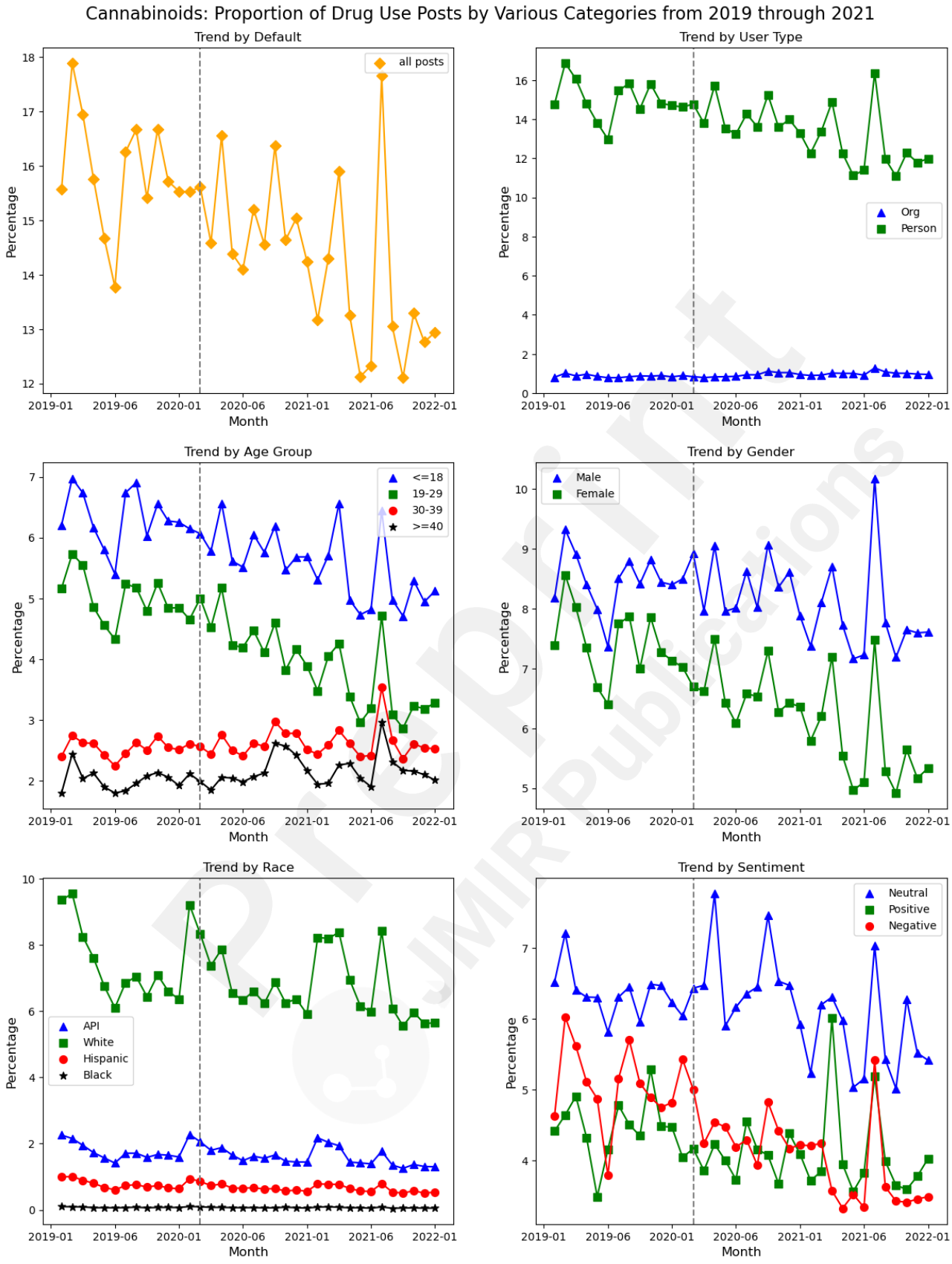
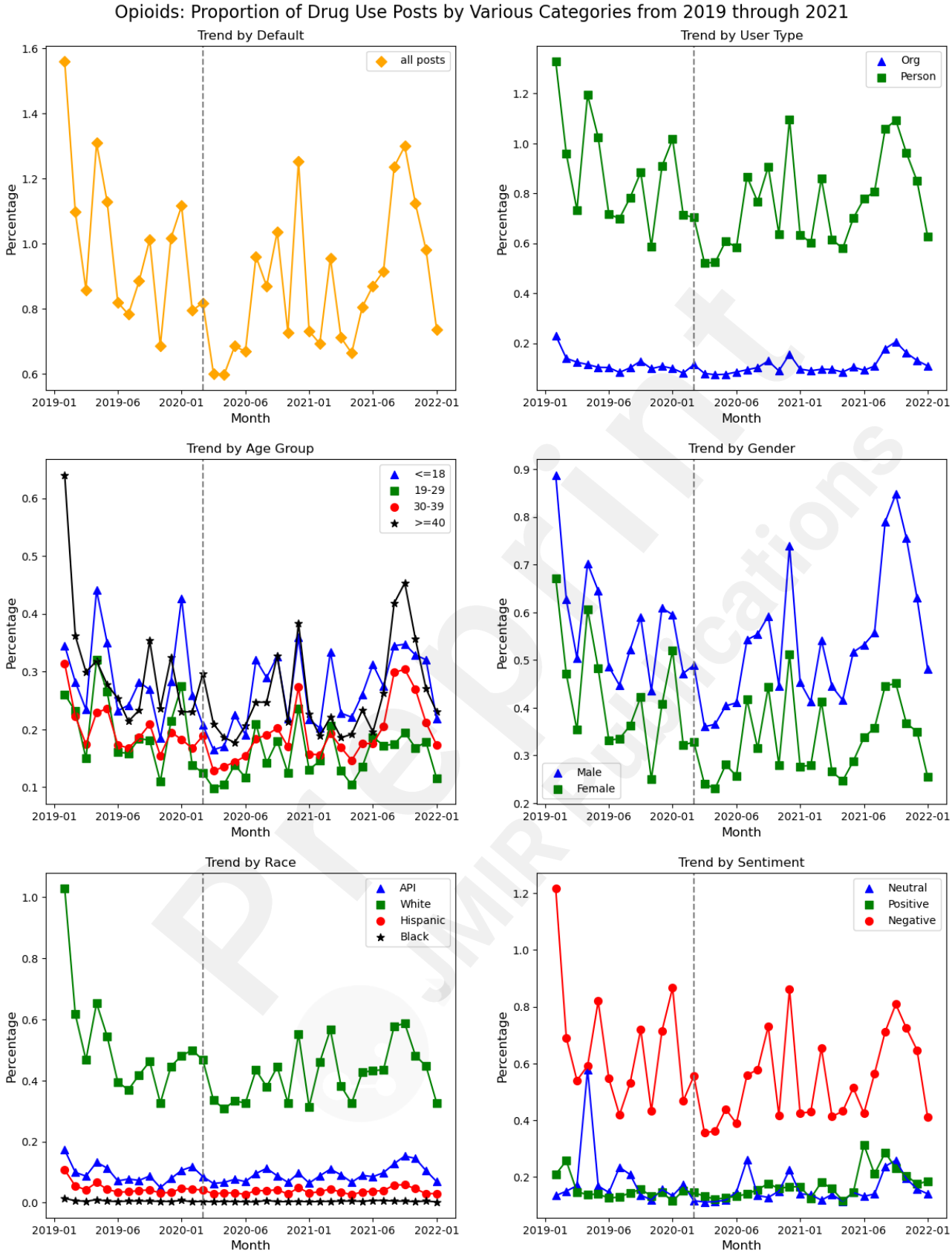


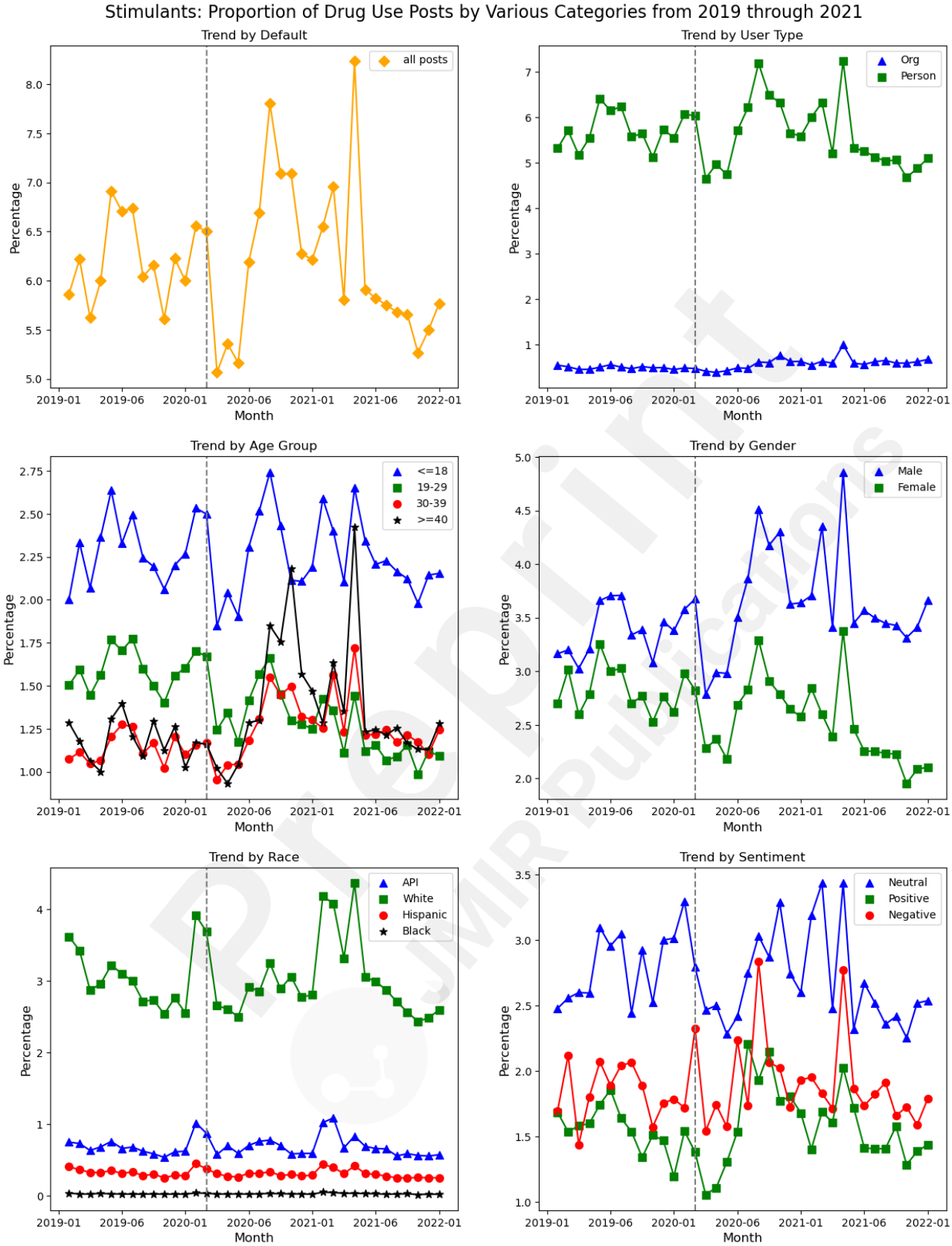
Figure 2. Trend in Substance Use Posts across demographic

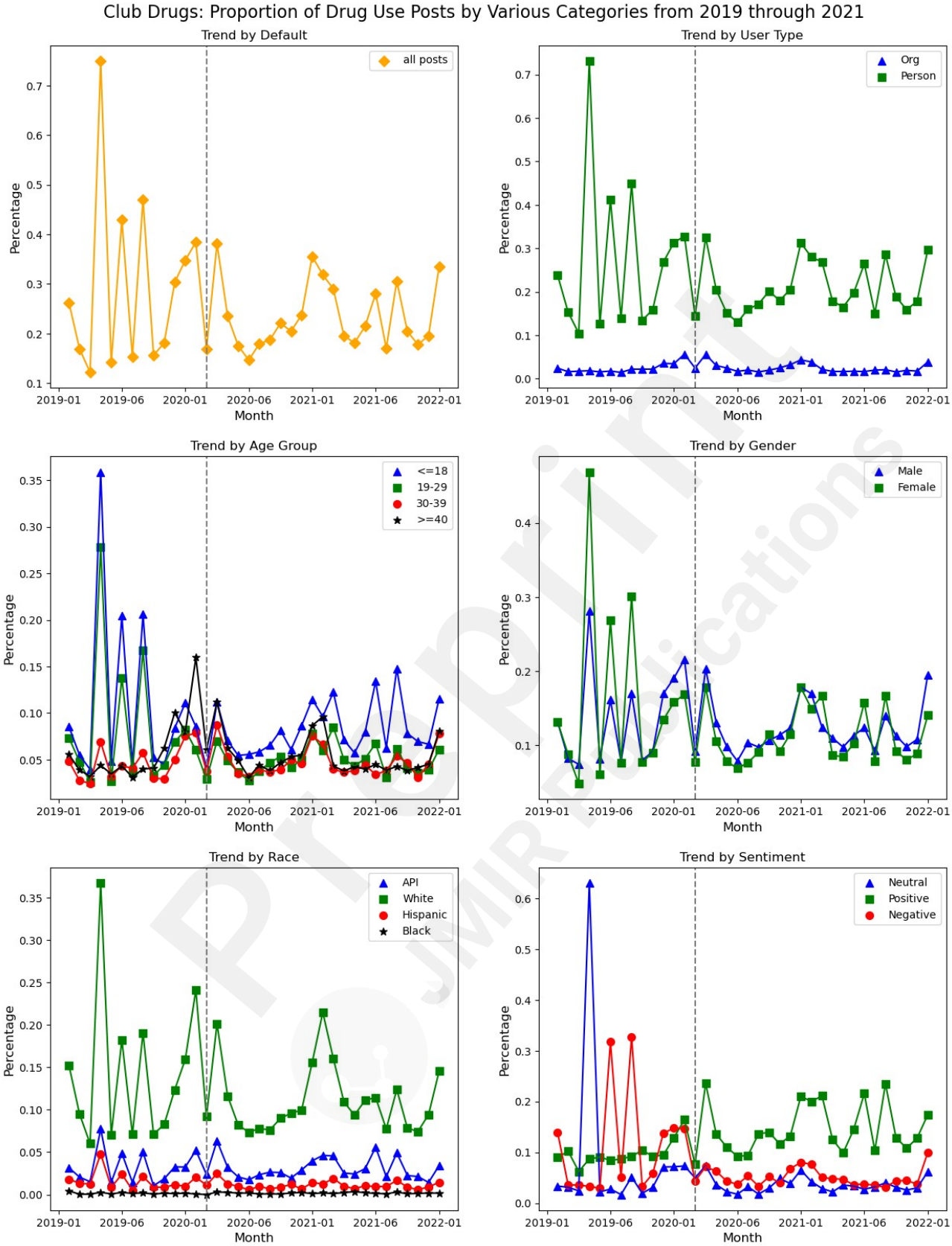


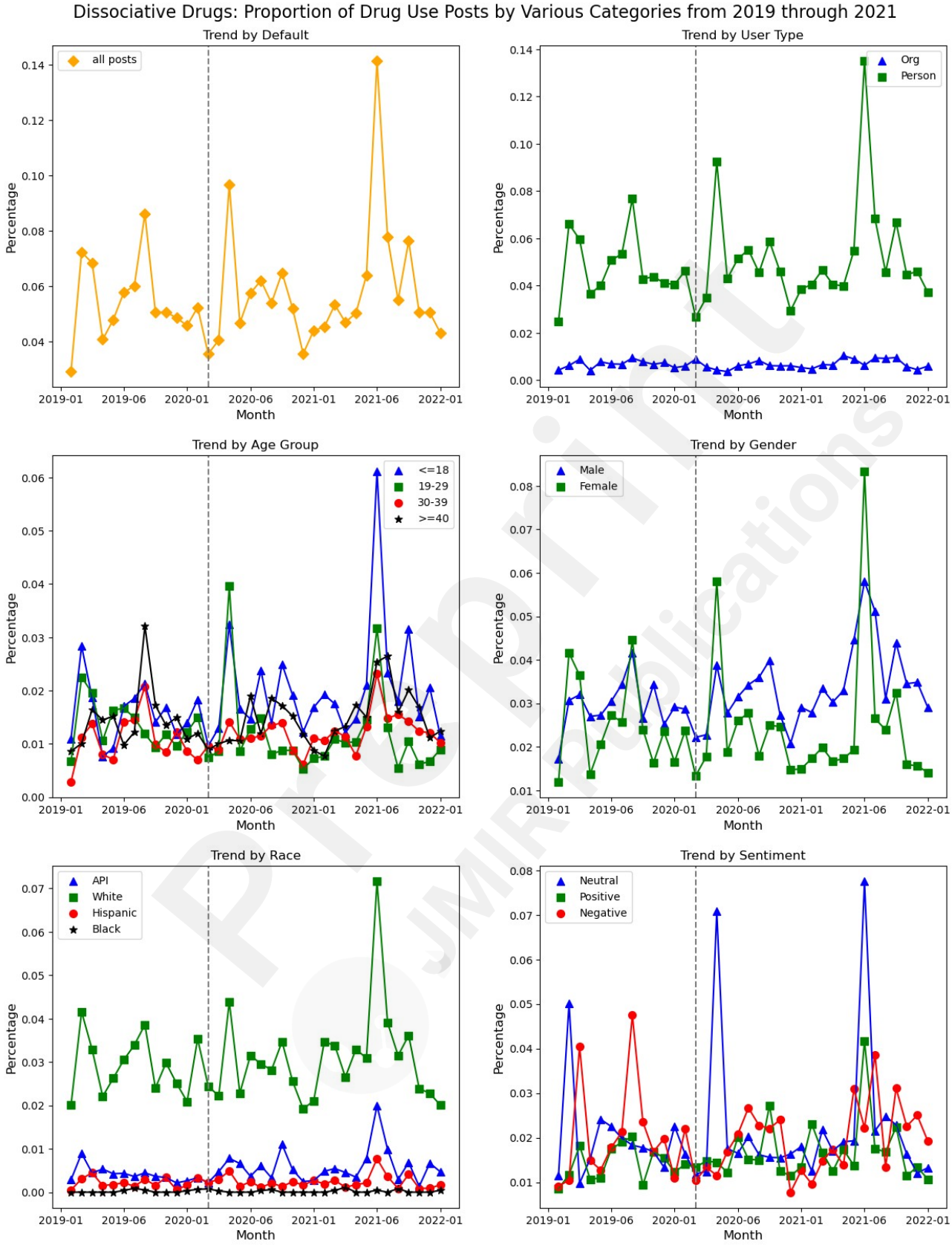


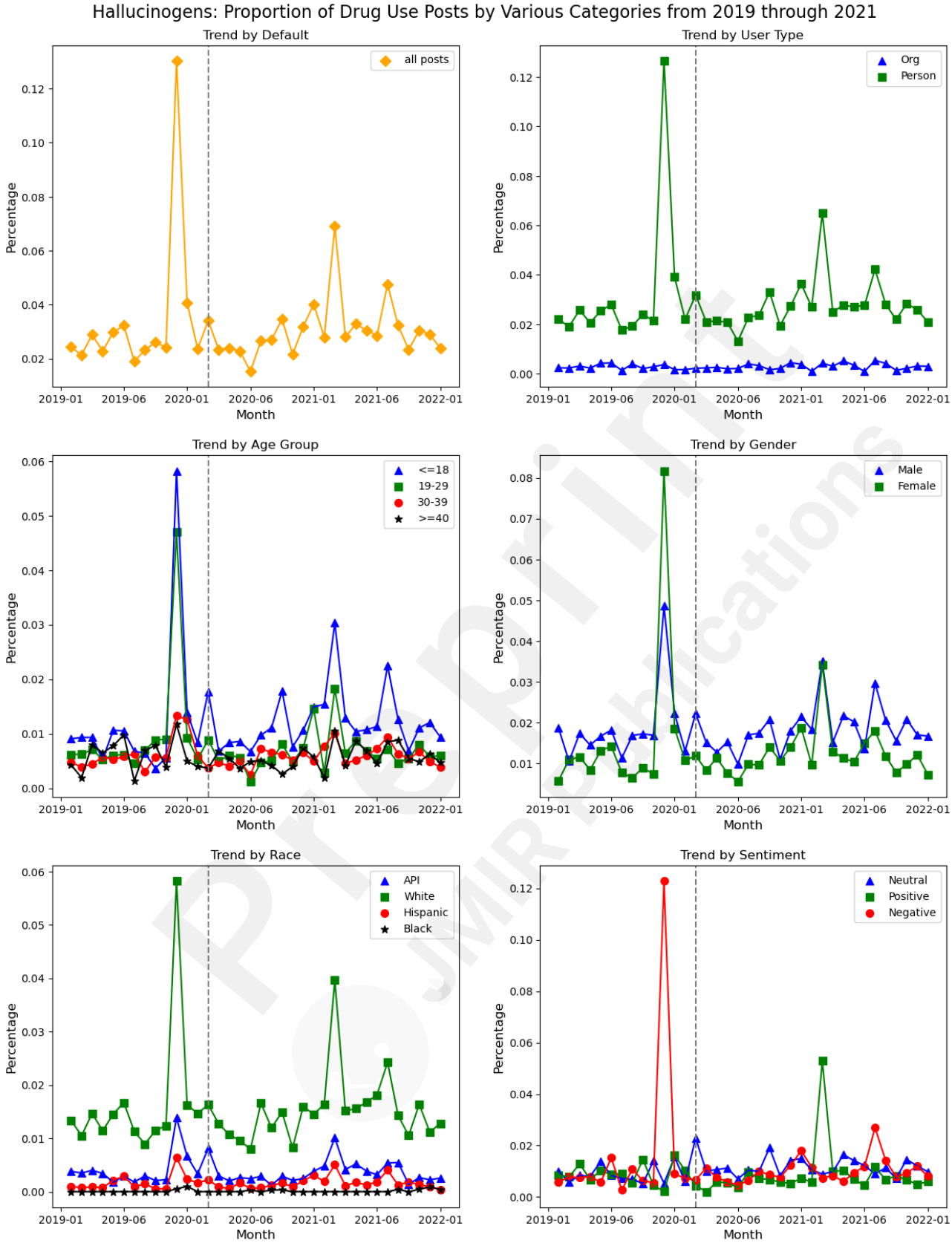

Cannabinoids: Proportion of Drug Use Posts by Various Categories from 2019 through 2021
Trend by Default
all posts
Trend by User Type
Org
Person
Trend by Age Group
<=18
19-29
30-39
>=40
Trend by Gender
Male
Female
Trend by Race
API
White
Hispanic
Black
Trend by Sentiment
Neutral
Positive
Negative
Percentage
Month
2019-01
2019-06
2020-01
2020-06
2021-01
2021-06
2022-01

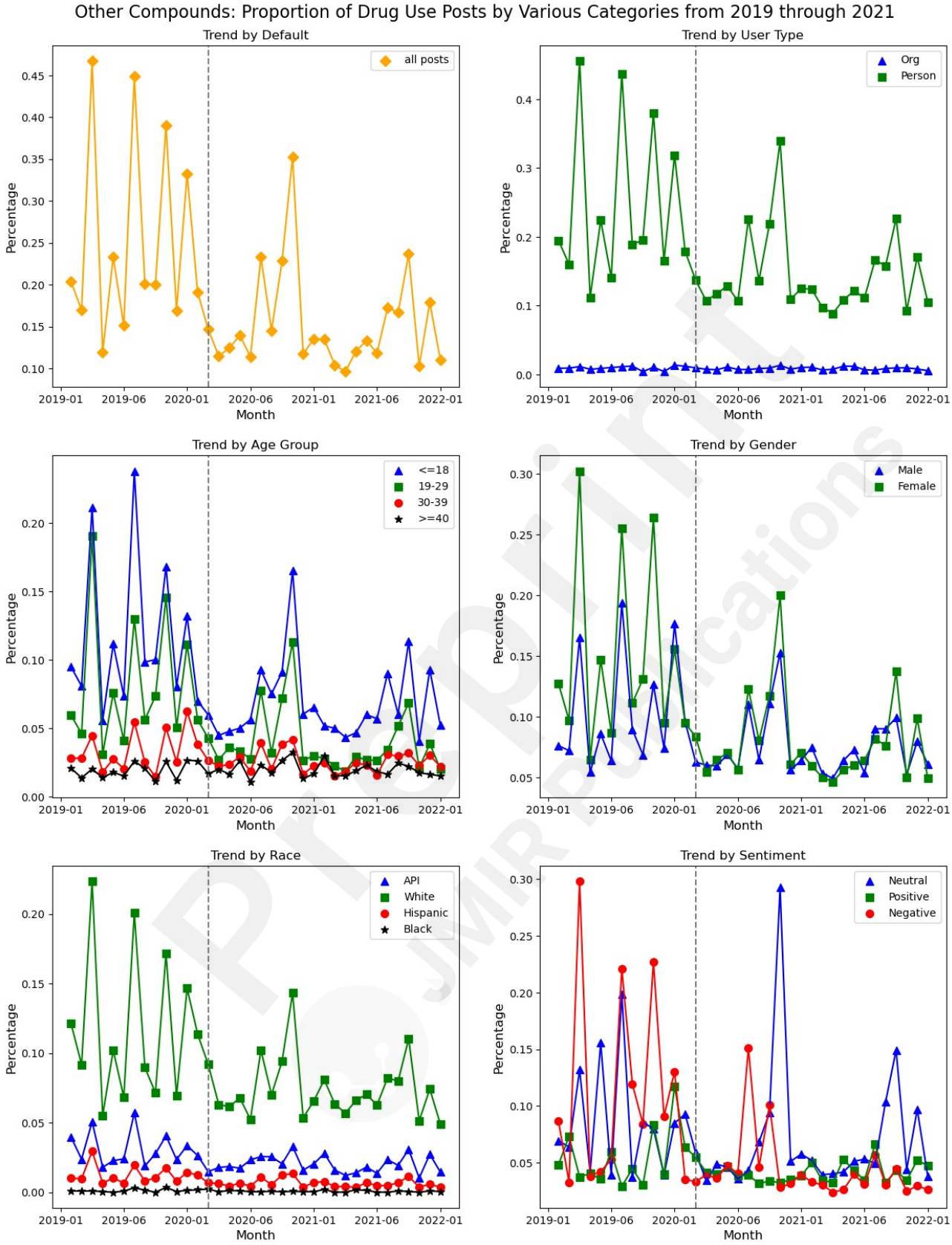


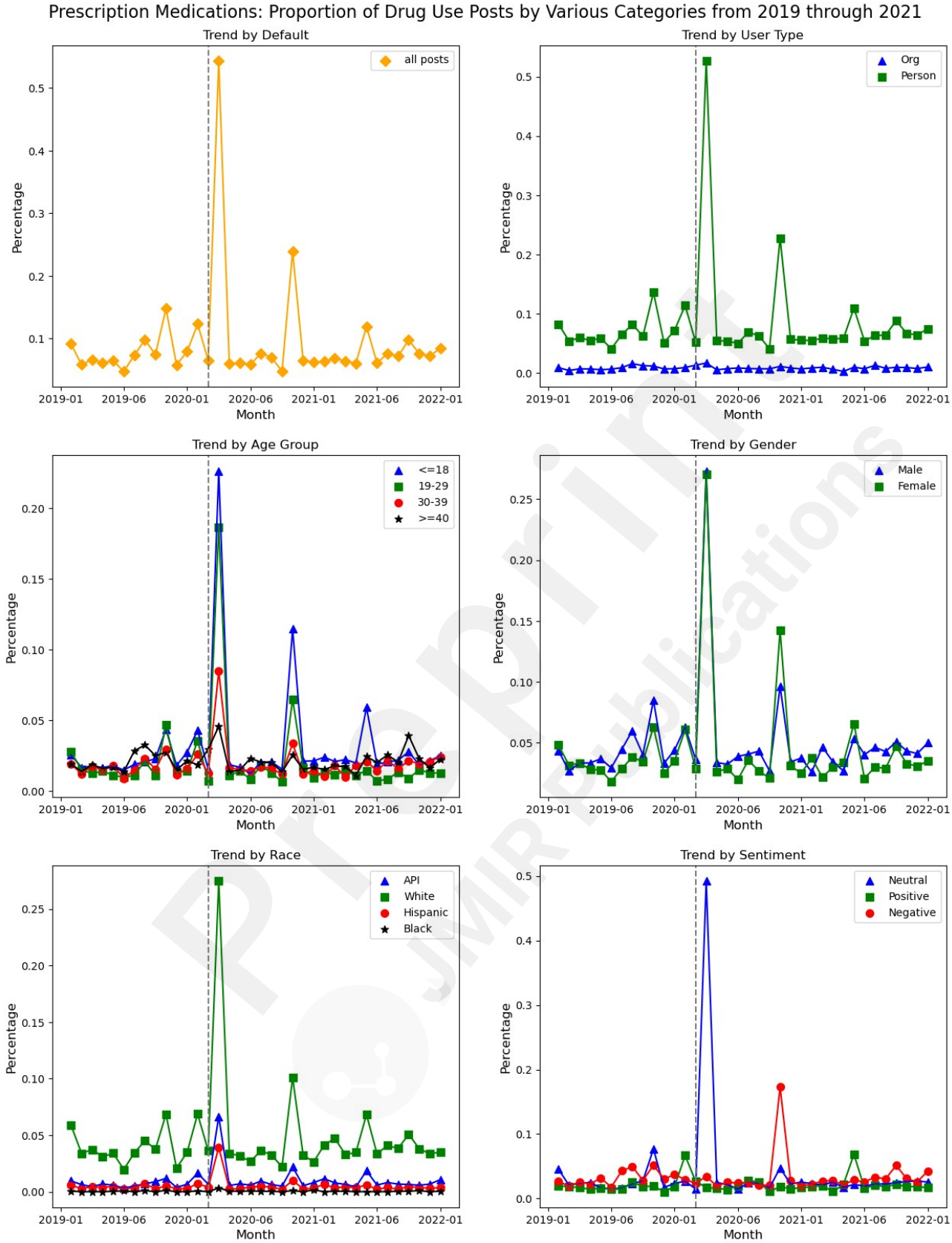











Prescription Medications: Proportion of Drug Use Posts by Various Categories from 2019 through 2021
Trend by Default
all posts
Trend by User Type
Org
Person
Trend by Age Group
<=18
19-29
30-39
>=40
Trend by Gender
Male
Female
Trend by Race
API
White
Hispanic
Black
Trend by Sentiment
Neutral
Positive
Negative
Percentage
Month
2019-01
2019-06
2020-01
2020-06
2021-01
2021-06
2022-01