Proceedings of the Nineteenth International AAAI Conference on Web and Social Media (ICWSM 2025)

Cannabis Use During Pregnancy: Insights from Online Discourse and Socioeconomic Indicators Across the USA and Canada

Lisette Espín-Noboa[1,2], Nikou Farsiu[3], Márton Karsai[1,4], Daniel J. Corsi[5]
[1]Department of Network and Data Science, Central European University, Vienna, Austria
[2]Complexity Science Hub, Vienna, Austria
[3]University of Tehran, Tehran, Iran
[4] National Laboratory for Health Security, HUN-REN Alfréd Rényi Institute of Mathematics, Budapest, Hungary
[5]Department of Obstetrics and Gynecology, Faculty of Medicine, University of Ottawa, Ottawa, Canada
espin@csh.ac.at, nikoufarsiu@gmail.com, karsaim@ceu.edu, dcorsi@ohri.ca

Abstract

Cannabis use is on the rise, driven by relaxing legal regulations and declining perceptions of harm. This trend, coupled with the increasing reliance on social media for health-related information, has sparked interest in cannabis use during pregnancy (**CanPreg**). This study examines online discourse about CanPreg on Twitter, analyzing 53,183 unique tweets from 32,744 users in the USA and Canada between 2012 and 2021. We investigate the spatio-temporal distribution of CanPreg discussions, key topical contexts within these conversations, and their correlations with socioeconomic and health indicators. The analysis reveals regional differences, with a relatively higher interest in CanPreg discussions in Canada compared to the USA. The online discourse is primarily focused on research, alongside criticism, personal experiences, queries, news sharing, and advertisements. Additionally, correlations between CanPreg tweet activity, poverty rates, and mental health metrics suggest a connection between online discussions and real-world behaviors. This study highlights the role of social media in health communication and provides insights to inform targeted intervention strategies.

Code — https://github.com/lisette-espin/CanPreg
Datasets — https://doi.org/10.17605/OSF.IO/P5D9Y

Introduction

Problem. The increasing prevalence of cannabis use during pregnancy (CanPreg) can be linked to factors such as more lenient legal regulations, shifting cultural perceptions, and the impact of the COVID-19 pandemic (Frank and Morrison 2022). Despite its widespread use, this prevalence remains understudied, leading to likely underestimation in reported exposure rates (Corsi et al. 2020). An increase in CanPreg is partly due to women using it to manage symptoms like nausea and appetite changes, believing it to be a natural and safer alternative compared to other substances (Chang et al. 2019). The lack of clear communication from healthcare providers about the adverse effects of cannabis use during pregnancy further contributes to this trend, as it leads expectant mothers to seek information from less reliable sources like social media and retail cannabis stores, where its use is often normalized (Frank and Morrison 2022).

Research Gap. Previous research has explored the use of Twitter and other social media platforms for cannabis use surveillance (Allem, Escobedo, and Dharmapuri 2020), identifying emerging consumption patterns (Meacham, Paul, and Ramo 2018), and monitoring the promotion and advertisement of cannabis products (Peiper et al. 2017). Studies have also examined the dissemination of anecdotal stories and misinformation about cannabis through social media. For instance, Ishida et al. found that individuals using social media for health information were 31% more likely than others to believe that cannabis use during pregnancy is safe, and 56% more likely to endorse general misinformation about cannabis (Ishida et al. 2020). Pregnant individuals are increasingly turning to online platforms for health advice, especially on emerging topics such as cannabis use. While prior work has examined cannabis-related discourse on Twitter, no studies have analyzed the spatio-temporal dynamics of this discourse over a decade-long period, nor have they linked these dynamics to a broad range of health, socioeconomic, and demographic indicators.

Research Questions. In this study, we address this gap by investigating the following research questions:

RQ1. **Which regions discuss CanPreg on Twitter, and how does the timing of these discussions vary?** We investigate the spatio-temporal distribution of users engaging in CanPreg-related discussions. Additionally, we examine whether increases in these conversations correlate with cannabis legalization policies.

RQ2. **What are the main topics related to CanPreg discussed on Twitter?** We perform content analysis to identify recurring topical contexts and concerns expressed in CanPreg-related tweets, shedding light on public perceptions, knowledge gaps, and common narratives in the online discourse.

RQ3. **Does the proportion of CanPreg-related tweets correlate with socioeconomic and health indicators?** We explore potential associations between the prevalence of CanPreg-related tweets and key socioeconomic and health metrics, such as poverty rates, drug use statistics, and access to healthcare.

Contributions. This decade-long longitudinal study demonstrates the potential of Twitter data, particularly the number of tweets and users discussing CanPreg topics, as a robust tool for monitoring diverse socioeconomic and health indicators. These include metrics related to poverty, mental health, demographic trends such as gender and age, and behavioral conditions like smoking and alcohol use. To advance transparency and facilitate future studies, we have made our code (Espín-Noboa 2025) and data (Espín-Noboa et al. 2025) publicly available. Although our analysis centers on Twitter and predominantly English-language data, the methods are designed to be adaptable for analyzing other platforms and languages, enabling broader applications.

Related Work

Health Implications of CanPreg. Cannabis use during pregnancy has been associated with significant short- and long-term health risks for both mothers and their offspring. For mothers, cannabis consumption may increase the risk of anemia. For fetuses and newborns, prenatal or breastfeeding exposure to cannabis has been linked to lower birth weights, often necessitating admission to neonatal intensive care units (Ryan et al. 2018). Infants exposed to cannabis in utero may exhibit impaired executive functioning, including reduced memory capacity, as well as physiological and behavioral disturbances such as tremors, heightened startle responses, disrupted sleep, and attention deficits (Barbosa-Leiker et al. 2020). These challenges can persist into early childhood, with affected children being more prone to hyperactivity, impulsivity, and emotional or behavioral problems, including depression and delinquent behaviors during adolescence. The harmful consequences of prenatal cannabis exposure may extend into adulthood, with lasting effects such as impaired visuo-spatial memory and an increased likelihood of developing drug-seeking behaviors (Barbosa-Leiker et al. 2020). Despite these risks, inadequate communication between healthcare providers and patients about cannabis use during pregnancy often leads expectant mothers to seek information from unreliable sources, such as social media or retail cannabis stores, where its use is normalized but poorly understood (Frank and Morrison 2022).

Health and Social Media. Social media platforms, particularly Twitter, have emerged as valuable tools for public health research and monitoring (Moorhead et al. 2013; Althouse et al. 2015; Sinnenberg et al. 2017). However, their open nature also brings challenges, such as the rapid dissemination of health misinformation. This can contribute to vaccine hesitancy, promote unproven treatments, and spread misleading information about drug safety, posing significant public health risks (Do Nascimento et al. 2022; Kbaier et al. 2024). Although users are often aware of the prevalence of false health information, many struggle to evaluate its accuracy, emphasizing the need for targeted intervention strategies (Kbaier et al. 2024). Despite these challenges, social media platforms hold potential for enhancing health communication when utilized effectively by professionals. For instance, studies have demonstrated Twitter's use in mental health advocacy (Berry et al. 2017) and improving public engagement through national health campaigns (Van Draanen et al. 2019). Additionally, social media is increasingly used to explore public sentiment on sensitive women's health topics. Research on abortion-related discussions on Facebook has revealed links between public attitudes, infant mortality rates, and political affiliations (Aleksandric et al. 2024).

CanPreg in Social Media. Pregnant women often avoid discussing cannabis use with healthcare providers due to stigma and legal concerns, turning instead to anonymous online platforms such as forums for information and support (Lebron et al. 2022). To better understand perceptions, trends, and discussions surrounding cannabis use during pregnancy, researchers have increasingly analyzed data from social media platforms, including Facebook, Twitter, and online forums (Oram et al. 2018; Cresswell et al. 2022). Quantitative studies have provided significant insights into substance use during pregnancy through social media analysis. Oram et al. examined Facebook posts from 43 pregnant adolescents, noting that 70% mentioned substance use before pregnancy, which decreased to 56% during pregnancy, accompanied by a shift toward a more negative tone (Oram et al. 2018). Similarly, Dakkak et al. analyzed 550 tweets with cannabis-related keywords, uncovering mixed perspectives on the risks and benefits of marijuana use during pregnancy (Dakkak et al. 2018). Pang et al. collected 17,238 Twitter posts over 12 months, coding 1,000 tweets to identify themes such as safety during pregnancy (36%), postpartum safety (2.3%), and cannabis use for pregnancy-related symptoms (2.7%) (Pang et al. 2021). Additionally, Lebron et al. analyzed 151 posts and 1,260 comments on the "Ganja Mamas" forum hosted on `Whattoexpect.com`, revealing common concerns about drug testing and interactions with child protective services (Lebron et al. 2022).

While these studies provide valuable insights, they are often limited by shorter time-frames and narrow sets of keywords. In contrast, our study leverages Twitter—a highly suitable platform for analyzing cannabis-related discussions (Cavazos-Rehg et al. 2016)—over a ten-year period, employing a broader and more comprehensive set of keywords. Furthermore, our analysis is robust, as it integrates Twitter data with official socioeconomic and health indicators to assess the significance of the online discourse and its alignment with real-world trends. This approach provides a more nuanced and reliable understanding of cannabis use during pregnancy as a critical public health issue.

Methods

Data Collection

Survey Data. To complement our analysis of Twitter data, we incorporated relevant socioeconomic and health indicators from official sources in the USA and Canada. These indicators provide essential context for interpreting the online discourse surrounding CanPreg. Table 1 outlines the data categories collected, including health statistics, demographic profiles, and economic metrics. For the USA, socioeconomic data was sourced from the U.S. Census Bureau[1], while health-related indicators, such as abortion rates and mortality data, were obtained from the U.S. Centers for

[1]https://www.census.gov/data.html

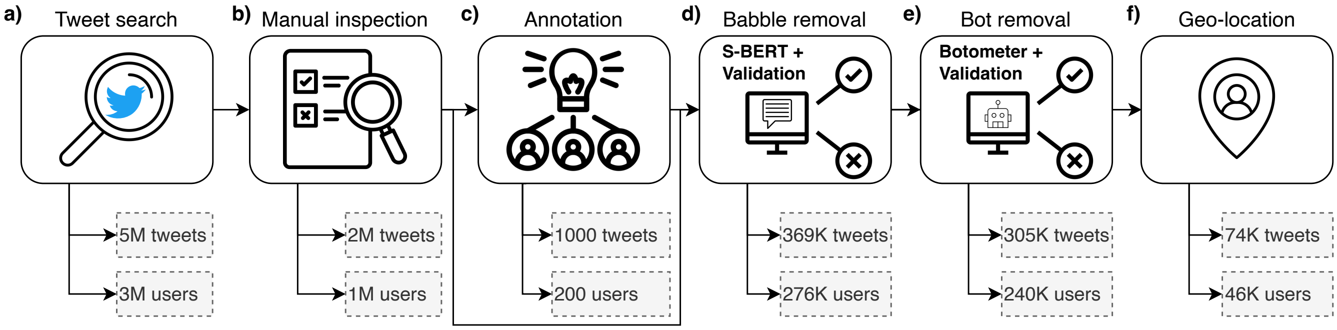


Figure 1: **Twitter data collection.** Tweets were collected during the period 2012–2021 using the Academic API v2.0, guided by two lists of keywords related to cannabis and pregnancy. To ensure data relevance, we manually inspected and removed irrelevant tweets (e.g., song lyrics). Automated filtering was applied using SBERT to exclude unrelated tweets and Botometer to identify and remove bot accounts. Each tweet was geo-located based on its tagged geo-location, the user's bio location, or the most frequently mentioned location in the user's tweet history.

#	Socio-economic-health indicator	USA	CAN
1	Abortions	✓	✓
2	General population	✓	✓
3	Population by gender	✓	✓
4	Population by age	✓	✓
5	Women in reproductive age	✓	✓
6	Health Insurance type	✓	✓
7	Mental health status	-	✓
8	Morbidity (deaths, live births)	✓	✓
9	Drug Abuse	-	✓
10	Poverty	✓	✓

Table 1: **Survey data.** Official survey data from the USA and Canada were collected for the period 2015–2021. The data includes information on abortions, population, drug abuse, poverty, and other relevant socioeconomic indicators at the state level for the USA and the provincial level for Canada.

Topic	Keywords
Cannabis (30)	cannabis, weed, pot, marijuana, marihuana, MJ, ganja, purp, bud, keef, kief, dope, "mary jane", thc, cbd, cannamom, opiate, mdma, ecstasy, mmj, medicalmarijuana, blunt, bong, budder, hash, hemp, indica, kush, reefer, sativa
Pregnancy (17)	pregnancy, pregnant, baby, fetus, fetal, prenatal, perinatal, womb, preggo, "pregnant life", "baby bump", "mom to be", "mommy to be", "baby on the way", "preggers", "pregnant af.", "pregnant as fuck"

Table 2: **Search keywords.** We use two independent lists of keywords: 30 related to cannabis and 17 related to pregnancy. These lists are combined using the Twitter API, applying the AND operator between the two groups and the OR operator within keywords of the same group. The selection of keywords builds upon prior work (Pang et al. 2021).

Disease Control and Prevention (CDC).[2] In Canada, most data were sourced from Statistics Canada[3] and information on induced abortions from (The Canadian Institute for Health Information (CIHI) 2025).

Twitter Data As illustrated in Figure 1.a, we first collected tweets potentially related to cannabis use during pregnancy (referred to as "CanPreg") posted from 2012-01-01 to 2021-12-31. We rely on the archival search functionality from the Twitter Academic API v2 using 30 cannabis- and 17 pregnancy-related keywords—extending the 25 keywords used in (Pang et al. 2021), see Table 2. Then, retweets, duplicates, and clearly non-related tweets were removed to focus on original content only (Figure 1.b).

Normalization. To enable fair comparisons across regions, we normalized all CanPreg counts by the total number of tweets (or unique users) posted within the same timeframes and regions.[4] Similarly, survey data were adjusted based on the population size of each region for each year, utilizing state-level data for the USA and provincial-level data for Canada. These normalization steps address group size biases, as larger regions naturally exhibit higher absolute Twitter activity. This bias is illustrated in Figure A8, which shows the relationship between region (population) size and Twitter activity in the data.

Pre-Processing

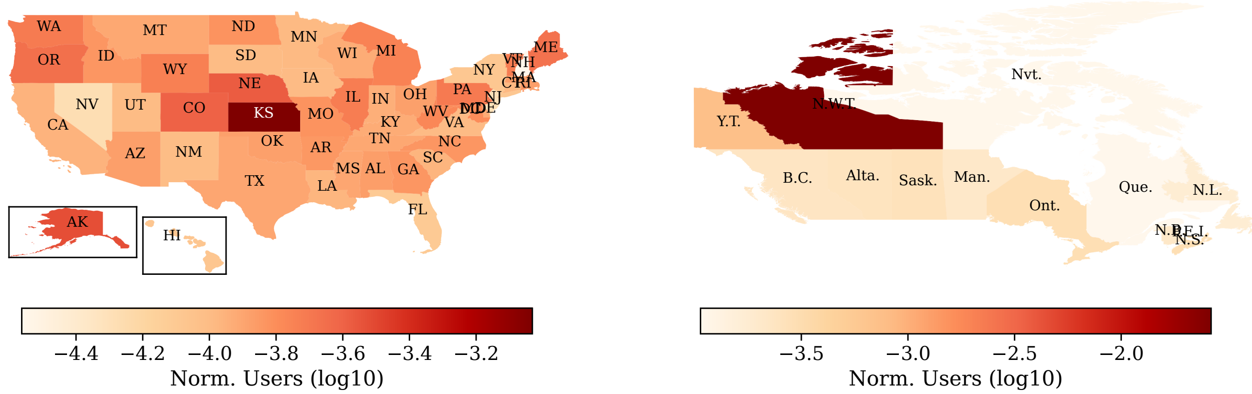
CanPreg Tweets. We automatized the removal of non-related tweets using SBERT (Reimers and Gurevych 2019b) semantic search (Figure 1.d). In a nutshell, we provided five queries[5] to S-BERT to get a semantic similarity score be-

[2]https://www.cdc.gov

[3]https://www.statcan.gc.ca

[4]Total tweet counts were only available from 2015-01-01.

[5]q_1: cannabis during pregnancy, q_2: smoking weed while pregnant, q_3: the effects of cannabis on pregnant women, q_4: smoking



(a) USA (b) Canada

Figure 2: **CanPreg users per region.** The figure presents the average normalized count of unique users posting CanPreg-related tweets, calculated as the yearly ratio of unique users to total users, averaged over 2015—2021. The left panel displays data for the USA, while the right panel focuses on Canada. In the USA, Kansas shows the highest proportion of CanPreg-active users annually, whereas the Northwest Territories lead in Canada. Counts are displayed on a base-10 logarithmic scale for clarity.

tween each tweet and query. We selected as valid all tweets that fulfill a minimum similarity score for at least one query (i.e., $q_1 = 0.43$, $q_2 = 0.48$, $q_3 = 0.42$, $q_4 = 0.42$, $q_5 = 0.43$), see "Validation sample annotation" for details.

CanPreg Users. Similarly, Botometer (Sayyadiharikandeh et al. 2020) was used to remove authors who are likely bot accounts (Figure 1.e). All accounts with a cap score[6] lower than or equal to 0.80 were selected as valid, see "Validation sample annotation" for details.

Validation Sample Annotation. To ensure the quality of our dataset and determine appropriate thresholds for SBERT and Botometer, we conducted a manual annotation process. We randomly sampled 1,000 tweets and 200 unique users from the collected data. Each sampled tweet was evaluated to determine whether it was genuinely related to CanPreg. Similarly, each author was assessed to determine whether they exhibited bot-like behavior. Each tweet and user was annotated by two independent researchers. In cases of disagreement, additional annotations were carried out by new researchers. More details about the annotations, including inter-annotator agreement and thresholds, are given in Table A1 and Figures A1 to A4 in the Appendix.

Geolocated Tweets. Geo-location data for authors was derived from their bio-descriptions, utilizing the lookup geometry functionality provided by Nominatim (Mocnik, Mobasheri, and Zipf 2018), or from the most frequent location associated with their CanPreg tweets. For tweets lacking a geo-tag, the author's location was used to infer their geo-location (Figure 1.f). The final dataset includes 73,891 unique tweets and 46,366 unique users globally.

or consuming drugs during pregnancy, q_5: the effects of cannabis on newborns.

[6]The cap score can be interpreted as the probability that an account with this score or greater is a bot.

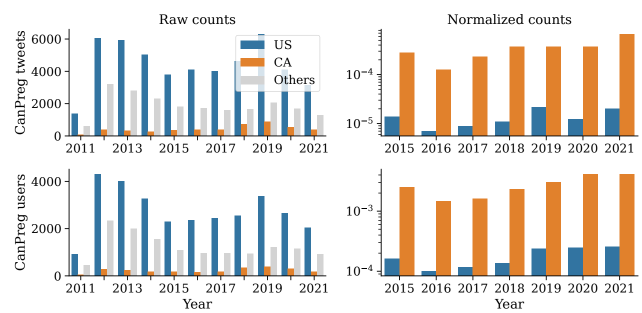


Figure 3: **CanPreg counts per year.** The left panel shows raw counts of CanPreg-related tweets (top) and unique users (bottom), while the right panel presents normalized counts, computed as the ratio of CanPreg-related tweets or users to total tweets or users per country (colors). Overall, the majority of tweets and users originate from the USA, with fewer from Canada. However, after normalization, the fraction of CanPreg-related tweets and users is relatively small in the USA but higher in Canada.

To address RQ1 and RQ2, we analyzed all valid data from the US and CA between 2012–2021. This dataset comprises 48,464 tweets (66% of the total) and 30,246 users (65%) from the US, and 4,719 tweets (6.4%) and 2,497 users (5.4%) from CA. For RQ3, we used data from 2015–2021, reflecting the availability of Twitter total counts.

Post-Processing

Clustering Tweets Based on Semantic Similarity. To group tweets with similar semantic meanings, we employed a clustering approach using the pre-trained Sentence-Transformers model `all-MiniLM-L6-v2` (Reimers and Gurevych 2019b). This model generates dense vector embeddings that capture the semantic content of sentences. For clustering, we use a fast and scalable algorithm optimized for large datasets (Reimers and Gurevych 2019a). It identifies communities—groups of embeddings that are closer to each other than a defined similarity threshold. Only clusters larger than a specified minimum size are returned, sorted by size, with each cluster's central point listed first. We configured the clustering algorithm with a minimum cluster size of 25 elements and a cosine similarity threshold of 0.75. These parameters ensure that each cluster contains a meaningful set of semantically related tweets, while avoiding overly broad or loosely connected groupings. To enhance interpretability, a domain expert manually labeled each cluster by assigning a general topic and a more specific subtopic. These labels were not predefined but were inductively derived by reading a random sample of 10 tweets from each cluster. The labeling followed consistent criteria and did not alter the clusters themselves.

Results

RQ1: Which Regions Discuss CanPreg on Twitter, and How Does the Timing of These Discussions Vary?

In Figure 3 (left), we see that the majority of CanPreg-related tweets and users originate from the USA, with fewer from Canada and other countries.[7] This disparity may be due to the English-only nature of our search keywords, despite retrieving tweets in any language, and the different level of popularity of the Twitter platform in the two countries. According to tweet metadata, 99.8% of the collected tweets are in English, although we identified 28 other languages during the initial data collection step (Figure 1.a).

However, when normalizing by the total number of tweets or users within the same time frame and country, Figure 3 (right) shows a higher proportion of CanPreg-related tweets and users in Canada than in the USA. This suggests that while the absolute counts is larger in the USA, the relative interest in CanPreg is stronger in Canada. Further, we observe a slight positive trend over time, particularly in Canada, indicating a growing public engagement with the topic. In Canada, cannabis has been legal for medicinal purposes since 2001 and was legalized for recreational use on October 17, 2018,[8] falling within our study period. In the USA, two distinct peaks are evident: one in 2012 and another in 2019. The 2012 peak may be linked to the legalization of marijuana use and sale in Colorado and Washington (Hickenlooper 2014), while the 2019 peak could be attributed to more than half of all state legislatures considering cannabis-related legislation during that year.[9]

Analyzing the normalized counts of unique users posting CanPreg-related tweets across regions (Figure 2), we identify some geographic patterns. In the USA, Kansas exhibits the highest proportion of CanPreg-active users relative to its overall tweeting population (Figure 2.a). Kansas's relatively small Twitter user base, comprising only 0.71% of all USA users, amplifies its normalized proportion, despite representing 4.85% of all CanPreg-related users. This may be partly explained by Kansas's historical context: between 2013 and 2015, the state considered legislation to ease marijuana restrictions.[10] Following Kansas, the District of Columbia, Alaska, Guam, Nebraska, and Colorado also rank among the top six regions for relative CanPreg activity.

In Canada, the Northwest Territories ranks as the top region for CanPreg-related user activity, accounting for 15% of all CanPreg users while representing only 0.16% of Canada's total active Twitter user base. It is followed by Yukon, another northern territory, along with Ontario, Prince Edward Island, Nova Scotia, and Saskatchewan. Except for Ontario, the other five regions are among Canada's least populated provinces and territories (Statistics Canada 2024) and have the lowest average Twitter activity. Ontario alone accounts for approximately 46% of both total active users and CanPreg-related users.

A detailed breakdown of raw and normalized tweet and user counts by region, aggregated and by year, is provided in Appendix Figures A9 to A12.

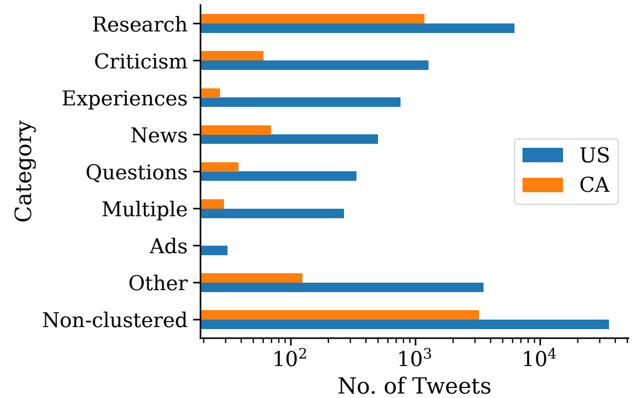


Figure 4: **Main CanPreg topics by country.** We identified 80 clusters using SBERT, which we manually labeled, resulting in six main topics: Research, Criticism, Experiences, News, Questions, and Ads. Clusters characterized by multiple main topics were combined under the label "Multiple." Non-clustered tweets refer to those that did not meet the thresholds of the clustering algorithm. In both countries, the majority of tweets were non-clustered. Among those with topical information, most tweets in both countries focused on sharing or commenting on research about CanPreg. In the USA, this was followed by tweets criticizing CanPreg and sharing personal experiences or opinions, while in Canada, news sharing and criticism ranked next. Advertisement content was uniquely found in the USA.

[7] The complete geographical distribution of CanPreg users worldwide is shown in Appendix Figure A5.

[8] https://www.justice.gc.ca/eng/cj-jp/cannabis

[9] https://thehill.com/homenews/state-watch/454167-record-number-of-states-considered-marijuana-legalization-in-2019

[10] https://www.washingtontimes.com/news/2015/may/26/kansas-marijuana-bill-hits-roadblock-in-state-sena

Cluster	# Tweets	Topic	Sub-topic	Example (paraphrased tweet)
1	6802	Research	CanPreg studies	Marijuana use doubles the risk of premature birth. [URL] via [MENTION]
3	532	Criticism	Criticism of CanPreg in general	I don't understand why pregnant women smoke weed or cigarettes smh.
5	316	Research	Baby boomers and cannabis use	Marijuana use is increasing among American baby boomers. [URL]
7	213	Questions	CBD safety during pregnancy	Is using CBD to manage pregnancy symptoms safe? [URL]
9	173	News	Maternal neglect with cannabis	[MENTION]: Mom drives off with baby on car roof after smoking pot. [URL]
...	...	...	...	...
42	41	Criticism	Smoking near infants	I don't support smoking weed around a baby.
43	41	Experiences	Cannabis and sleep benefits	That weed made me so sleepy—I slept like a baby.
44	39	Experiences, Research	CanPreg and breastfeeding	THC in breastmilk may cause drowsiness, but I didn't notice this with my baby.
45	37	Experiences	Morning sickness	Don't want to smoke pregnant, but it eases nausea. #life
46	37	Criticism	Smoking near infants	Why smoke weed with a baby in the car? Seriously.
...	...	...	...	...
72	27	News	Cannabis warning labels for pregnancy	New bill: marijuana warnings for pregnant women. [URL]
75	26	Criticism	Criticism of CanPreg in general	I support smoking weed, but doing it while pregnant is just not right, smh.
76	26	Questions	CanPreg populaiton	Has anyone smoked weed during pregnancy?
77	26	News	Smuggling marihuana as baby bump	Woman caught with 'baby bump' hiding marijuana. [URL]
80	25	News	Criticism of CanPreg in general	Surgeon General warns teens, pregnant women: no marijuana. [URL]
(Multi)	3612	Others	Lyrics, babble, non-CanPreg	Haven't smoked weed in 6 months. That first one will hit hard! [URL]
(None)	38819	Un-clustered	Various	Light up before intimacy—it's the trick to finishing without consequences. #magic

Table 3: **Semantic clusters subset.** Tweets were grouped into 80 semantically related clusters using a Sentence Transformer model. This table showcases the top 5, middle 5, and bottom 5 clusters based on tweet counts across both countries. Each cluster is manually assigned a general *topic* and a specific *sub-topic* for clarity. Examples are randomly selected and paraphrased tweets. We grouped 24 clusters as "Others" for babble or non-CanPreg tweets. All clusters are detailed in Appendix Table A6.

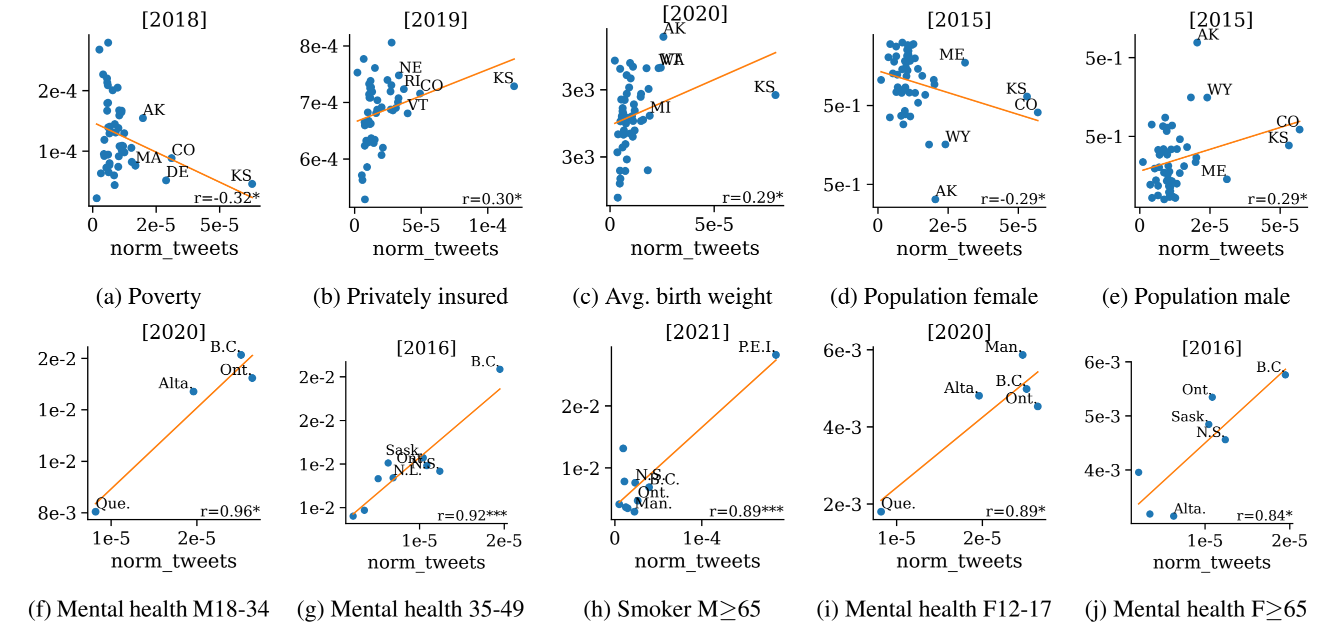
RQ2: What Are the Main Topics Related to CanPreg Discussed on Twitter?

Using a fast clustering algorithm (Reimers and Gurevych 2019a) based on Sentence Transformers (Reimers and Gurevych 2019b), we identified 80 clusters, grouping tweets based on semantic similarity. Then, each cluster was manually labeled, resulting in 41 distinct sub-topics grouped into six main topics (see "Methods" for details). Clusters may fit multiple main topics (e.g., cluster #44), while tweets containing lyrics or nonspecific content were categorized as "Other," encompassing 24 clusters. Non-clustered tweets are those that did not meet the clustering thresholds (i.e., fewer than 25 tweets or cosine similarity below 0.75). All clusters are detailed in Appendix Table A6.

Table 3 highlights 15 clusters: the top 5, middle 5, and bottom 5 by tweet count. Figure 4 illustrates the distribution of tweets per cluster topic across both countries. Among those with topical context, Research is the most common topic in both countries, with users discussing or sharing studies and sources related to CanPreg.

In the USA (Figure 4, blue), the second most frequent topic is Criticism, targeting CanPreg practices broadly or pointing out specific individuals (e.g., acquaintances or relatives). This is followed by personal Experiences, where users discuss whether CanPreg has been beneficial to themselves or someone they know. Americans also share News, including cases of medical marijuana saving infants, legislation updates, and incidents of poor parenting linked to marijuana misuse. Additional discussions include Questions about CanPreg's effects on mothers, fetuses, and even fathers (e.g., its impact on sperm count). Advertisements, though less common, were found exclusively in the USA.

In Canada (Figure 4, orange), the topic distribution dif-



(a) Poverty (b) Privately insured (c) Avg. birth weight (d) Population female (e) Population male

(f) Mental health M18-34 (g) Mental health 35-49 (h) Smoker M≥65 (i) Mental health F12-17 (j) Mental health F≥65

Figure 5: **Top-5 correlations between survey data and CanPreg tweets.** The strongest correlations for the USA (top) and Canada (bottom) are shown across years (columns). In the USA, negative correlations are observed between CanPreg tweet proportions and the fraction of individuals in poverty and women, while positive correlations are found with average newborn weight, the proportion of privately insured people, and men. In Canada, higher prevalence of poor mental health and elder smokers correlates positively with CanPreg tweet proportions. All correlations are statistically significant ($p \leq 0.05$).

fers. After Research, News is the second most popular topic, followed by Criticism, Questions, and personal Experiences.

Figure A6 and Figure A7 in the Appendix illustrate, respectively, the frequency of tweets containing each keyword used in our Twitter search, and a word cloud of non-keyword-related terms found in valid tweets. Among the 47 keywords employed, 45 appeared in the USA and 43 in Canada within the valid geo-located dataset. In Canada, "cannabis" and "pregnant" were the most frequently used keywords, while in the USA, "weed" and "baby" were predominant. The ambiguity of the word "baby" led to many unrelated tweets referencing song lyrics, slang, or general chatter. This accounts for the reduction from 5M to 369K tweets after filtering unrelated content, as shown in Figure 1.a and Figure 1.e. The word clouds in Appendix Figure A7 reveal that sharing URLs, mentions of "women" and "smoking" are prevalent in tweets from both countries.

RQ3: Does the Proportion of CanPreg-Related Tweets Correlate With Socioeconomic and Health Indicators?

Figure 5 presents the top five strongest and statistically significant correlations[11] between survey data and *tweet* activity per year in the USA (top panel) and Canada (bottom panel). Comprehensive details on these correlations are provided in Appendix Tables A3 and A5.

In the USA, the proportion of CanPreg-related tweets shows positive correlations with some variables in specific years. These include the average weight of newborns in 2020, the size of population privately insured in 2019, and the population of men in 2015. Conversely, negative correlations are evident with the population in poverty in 2018 and the female population in 2015. These results suggest that states with higher poverty levels and larger female populations tend to have fewer CanPreg discussions on Twitter.

In Canada, correlations derived from more granular surveys reveal nuanced patterns. In 2015, CanPreg tweet activity correlates positively with poor mental health among adult females and adult males who are heavy drinkers. In 2016 and 2018, CanPreg tweet activity shows a positive correlation with mental health challenges, especially among young adults, reflecting heightened discussions in regions experiencing greater mental health issues. By 2019, a shift occurs, with negative correlations observed between CanPreg tweet activity and smoking prevalence, particularly among adults and the elderly. This suggests that regions with higher smoking rates engage less in CanPreg-related discussions. In 2020, tweet activity ratios indicate a strong positive correlation with poor mental health, predominantly among female teenagers and adult males, as well as among low-income individuals and those who have access to a regular health care provider. Conversely, better mental health in elderly and female populations aligns with reduced engagement in

[11]We computed Pearson correlation coefficients and associated two-sided p-values using the `pearsonr` function from the `scipy.stats` module in Python. This test evaluates the null hypothesis of zero linear correlation between two variables, under the assumption that both variables are normally distributed.

CanPreg discussions. In 2021, patterns diversify, with correlations involving behaviors like smoking, heavy drinking, and daily cannabis use. Regions with higher daily cannabis use among adults and the elderly exhibit increased CanPreg-related activity on Twitter.

Significant correlations between survey data and *users* are detailed in Appendix Tables A2 and A4, with the five strongest correlations summarized in Appendix Figure A13.

Limitations

Our study has some limitations that should be considered when interpreting the findings.

Access to Twitter Data. We collected data in 2022 while free academic access to the Twitter API v2.0 was still available. Since February 9th, 2023, this API has required a paid subscription,[12] limiting future use of the proposed methodology unless the data is purchased. Despite this constraint, our findings offer valuable insights in two ways. First, they suggest that online discourse, particularly on Twitter, can act as a proxy for socioeconomic and health indicators, enabling periodic monitoring to inform healthcare and policymaking efforts. Second, this highlights the potential of online discourse analysis on other platforms, such as Mastodon, Bluesky, Reddit, and comment sections on platforms like YouTube and TikTok, to develop monitoring tools adapted to the diverse audiences of these services.

English-Only Keywords. Although we did not explicitly constrain the language of tweets, we only provided English keywords to the Twitter API. This limitation restricts a broader understanding of CanPreg-related discourse in non-English-speaking regions. Future research could target specific countries by adapting these keywords to the languages spoken in those regions or by expanding the keyword lists to better capture the local context. Nevertheless, we make our code and data available for replication and extension in other countries, including the UK, Australia, and additional regions where relevant comparisons can be drawn.

Causality. The correlations observed in RQ3 reflect associations, not causal links, and should not be misused to draw unwarranted conclusions about the influence of CanPreg activity on societal indicators or vice versa. Responsible application of these insights is essential to avoid oversimplifying or misrepresenting these nuanced dynamics.

Scalability of Cluster Labeling. Given the small number of clusters identified in our study, manual interpretation was feasible and appropriate. However, future analyses on larger datasets may benefit from fully or partially automating this process. Recent advances provide sophisticated methods for extracting interpretable, high-level concepts from unstructured text (Pacheco et al. 2023; Lam et al. 2024).

Conclusion

Cannabis use, particularly during pregnancy, is rising, yet its effects on mothers and babies remain poorly understood due to the limited and inconclusive state of medical research. Pregnant individuals often turn to online platforms, including social media, to seek information on cannabis use during pregnancy (CanPreg). While these channels offer an open avenue for discussion, they frequently lack reliability. Healthcare professionals face challenges in gathering accurate data on cannabis use due to stigma, privacy concerns, and the sensitive nature of the topic. This creates a critical need for alternative data sources to better understand societal behaviors and attitudes around CanPreg.

In this longitudinal study, we leveraged ten years of Twitter data (2012–2021) to analyze discussions about CanPreg in the USA and Canada. By applying a robust methodology to exclude bots and unrelated content, we identified CanPreg-related tweets and mapped them to spatio-temporal dimensions (RQ1). We further explored the topical context of these conversations (RQ2) and assessed whether CanPreg activity on Twitter serves as a proxy for socioeconomic and health indicators (RQ3). Our focus on the USA and Canada, the first and third largest sources of CanPreg-related content, allowed us to better understand this online phenomenon.

Our findings show that CanPreg-related discussions occur across both the USA and Canada. Although the number of tweets is higher in the USA, Canadian users exhibit relatively greater engagement with CanPreg, suggesting the topic holds particular significance in Canada. This heightened interest may partly stem from cannabis legalization in Canada in 2018, which falls within the study period. In relative terms, low-density regions such as Kansas in the USA and the Northwest Territories in Canada exhibit a disproportionately high focus on CanPreg relative to their total Twitter activity. Overall, the discussions encompass a range of topics, including research findings, opinions on maternal cannabis use, personal experiences, news coverage, and questions about its effects. Most tweets express critical or cautious perspectives, particularly regarding cannabis use during pregnancy and its potential impact on newborns, though a smaller subset conveys anecdotal support.

Importantly, we found that CanPreg activity on Twitter statistically correlates with offline socioeconomic and health indicators. In the USA, CanPreg signals align with poverty rates, private insurance coverage, and female population size. In Canada, CanPreg discussions are linked to mental health challenges and alcohol abuse, highlighting its potential as a proxy for these complex societal issues. However, these findings should be interpreted responsibly. While our analysis identifies strong and statistically significant relationships, it does not establish causality or directionality.

While our study focuses on the USA and Canada, we share our code (Espín-Noboa 2025) and data (Espín-Noboa et al. 2025) to foster replicability and encourage future research in other regions or contexts. Although Twitter's changing policies may limit researchers' ability to freely access data on this platform, the insights from this work underline the value of using alternative online platforms as monitoring systems for sensitive health issues. By capturing public discourse, this study not only contributes to understanding societal attitudes and behaviors but also provides valuable tools to healthcare professionals and policymak-

[12]https://www.forbes.com/sites/jenaebarnes/2023/02/03/twitter-ends-its-free-api-heres-who-will-be-affected

ers to monitor trends and identify at-risk populations. The findings emphasize the importance of interdisciplinary approaches integrating computational social science with public health to address emerging societal challenges.

Acknowledgments

We are grateful to Thomas Louf for providing the total counts of unique tweets and active users on Twitter per region from 2015–2021. Additionally, we extend our sincere thanks to Adriana Manna, Elsa Andres, Hao Cui, Iacopo Iacopini, Júlia Számely, Katherine Muldoon, Kimberly Grattan, Ludovico Napoli, Malia Murphy, Remi Vaudaine, Sara Scremin, Serine Ramlawi, Sicheng Dai, and Stephanie Boyd for their valuable assistance with the annotation of tweets and users in our random sample, which was essential for creating the ground truth for this study. The authors would also like to thank Roberto Ulloa for his support with the Twitter Academic Application Programming Interface, and Indira Sen for her advice on the state-of-the-art natural language processing techniques. L.E.N. received support from the Vienna Science and Technology Fund WWTF under project No. ICT20-079. M.K. acknowledges funding from the National Laboratory for Health Security (RRF-2.3.1-21-2022-00006); SoBigData PPP: HORIZON-INFRA-2021-DEV-02 project (101079043); and the COLINE DUT FFG projects. This work was supported by a Canadian Institutes of Health Research Team Grant awarded to D.J.C. (funding reference CA3-170126).

References

Aleksandric, A.; Anderson, H. I.; Dangal, A.; Wilson, G. M.; and Nilizadeh, S. 2024. Analyzing the Stance of Facebook Posts on Abortion Considering State-Level Health and Social Compositions. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 15–28.

Allem, J.-P.; Escobedo, P.; and Dharmapuri, L. 2020. Cannabis surveillance with Twitter data: emerging topics and social bots. *American journal of public health*, 110(3): 357–362.

Althouse, B. M.; Scarpino, S. V.; Meyers, L. A.; Ayers, J. W.; Bargsten, M.; Baumbach, J.; Brownstein, J. S.; Castro, L.; Clapham, H.; Cummings, D. A.; et al. 2015. Enhancing disease surveillance with novel data streams: challenges and opportunities. *EPJ data science*, 4(1): 1–8.

Barbosa-Leiker, C.; Burduli, E.; Smith, C. L.; Brooks, O.; Orr, M.; and Gartstein, M. 2020. Daily cannabis use during pregnancy and postpartum in a state with legalized recreational cannabis. *Journal of addiction medicine*, 14(6): 467–474.

Berry, N.; Lobban, F.; Belousov, M.; Emsley, R.; Nenadic, G.; and Bucci, S. 2017. #WhyWeTweetMH: understanding why people use Twitter to discuss mental health problems. *Journal of medical Internet research*, 19(4): e107.

Cavazos-Rehg, P. A.; Sowles, S. J.; Krauss, M. J.; Agbonavbare, V.; Grucza, R.; and Bierut, L. 2016. A content analysis of tweets about high-potency marijuana. *Drug and alcohol dependence*, 166: 100–108.

Chang, J. C.; Tarr, J. A.; Holland, C. L.; De Genna, N. M.; Richardson, G. A.; Rodriguez, K. L.; Sheeder, J.; Kraemer, K. L.; Day, N. L.; Rubio, D.; et al. 2019. Beliefs and attitudes regarding prenatal marijuana use: perspectives of pregnant women who report use. *Drug and alcohol dependence*, 196: 14–20.

Corsi, D. J.; Donelle, J.; Sucha, E.; Hawken, S.; Hsu, H.; El-Chaâr, D.; Bisnaire, L.; Fell, D.; Wen, S. W.; and Walker, M. 2020. Maternal cannabis use in pregnancy and child neurodevelopmental outcomes. *Nature medicine*, 26(10): 1536–1540.

Cresswell, L.; Espín-Noboa, L.; Murphy, M. S.; Ramlawi, S.; Walker, M. C.; Karsai, M.; Corsi, D. J.; et al. 2022. The volume and tone of Twitter posts about cannabis use during pregnancy: protocol for a scoping review. *JMIR Research Protocols*, 11(3): e34421.

Dakkak, H.; Brown, R.; Twynstra, J.; Charbonneau, K.; and Seabrook, J. 2018. The perception of pre-and post-natal marijuana exposure on health outcomes: A content analysis of Twitter messages. *Journal of neonatal-perinatal medicine*, 11(4): 409–415.

Do Nascimento, I. J. B.; Pizarro, A. B.; Almeida, J. M.; Azzopardi-Muscat, N.; Gonçalves, M. A.; Björklund, M.; and Novillo-Ortiz, D. 2022. Infodemics and health misinformation: a systematic review of reviews. *Bulletin of the World Health Organization*, 100(9): 544.

Espín-Noboa, L. 2025. CanPreg. https://github.com/lisette-espin/CanPreg. GitHub repository.

Espín-Noboa, L.; Farsiu, N.; Corsi, D. J.; and Márton, K. 2025. CanPreg Datasets. https://doi.org/10.17605/OSF.IO/P5D9Y. OSF Dataset.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles. Accessed April 23, 2025.

Frank, C. J.; and Morrison, L. 2022. Cannabis Use During Pregnancy. *American Family Physician*, 106(4): 364–365.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Hickenlooper, G. J. W. 2014. Experimenting with pot: The state of Colorado's legalization of marijuana. *The Milbank Quarterly*, 92(2): 243.

Ishida, J. H.; Zhang, A. J.; Steigerwald, S.; Cohen, B. E.; Vali, M.; and Keyhani, S. 2020. Sources of information and beliefs about the health effects of marijuana. *Journal of general internal medicine*, 35: 153–159.

Kbaier, D.; Kane, A.; McJury, M.; and Kenny, I. 2024. Prevalence of health misinformation on social media—challenges and mitigation before, during, and beyond the covid-19 pandemic: Scoping literature review. *Journal of Medical Internet Research*, 26: e38786.

Lam, M. S.; Teoh, J.; Landay, J. A.; Heer, J.; and Bernstein, M. S. 2024. Concept induction: Analyzing unstructured text with high-level concepts using lloom. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–28.

Lebron, C. N.; Morales, V.; Saenz, S.; and Vidot, D. C. 2022. "Ganja Mamas": Online discussions about cannabis use in pregnancy. *Drug and alcohol dependence*, 241: 109689.

Meacham, M. C.; Paul, M. J.; and Ramo, D. E. 2018. Understanding emerging forms of cannabis use through an online cannabis community: an analysis of relative post volume and subjective highness ratings. *Drug and alcohol dependence*, 188: 364–369.

Mocnik, F.-B.; Mobasheri, A.; and Zipf, A. 2018. Open source data mining infrastructure for exploring and analysing OpenStreetMap. *Open Geospatial Data, Software and Standards*, 3(1): 1–15.

Moorhead, S. A.; Hazlett, D. E.; Harrison, L.; Carroll, J. K.; Irwin, A.; and Hoving, C. 2013. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. *Journal of medical Internet research*, 15(4): e1933.

Oram, D.; Wernette, G. T.; Nichols, L. P.; Vydiswaran, V. V.; Zhao, X.; Chang, T.; et al. 2018. Substance use among young mothers: an analysis of Facebook posts. *JMIR Pediatrics and Parenting*, 1(2): e10261.

Pacheco, M. L.; Islam, T.; Ungar, L.; Yin, M.; and Goldwasser, D. 2023. Interactive Concept Learning for Uncovering Latent Themes in Large Text Collections. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Pang, R. D.; Dormanesh, A.; Hoang, Y.; Chu, M.; and Allem, J.-P. 2021. Twitter posts about cannabis use during pregnancy and postpartum: a content analysis. *Substance Use & Misuse*, 56(7): 1074–1077.

Peiper, N. C.; Baumgartner, P. M.; Chew, R. F.; Hsieh, Y. P.; Bieler, G. S.; Bobashev, G. V.; Siege, C.; and Zarkin, G. A. 2017. Patterns of Twitter behavior among networks of cannabis dispensaries in California. *Journal of medical Internet research*, 19(7): e236.

Reimers, N.; and Gurevych, I. 2019a. Clustering with Sentence Transformers. https://www.sbert.net/examples/sentence_transformer/applications/clustering/. Accessed: 2025-04-04.

Reimers, N.; and Gurevych, I. 2019b. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ryan, S. A.; Ammerman, S. D.; O'Connor, M. E.; Gonzalez, L.; Patrick, S. W.; Quigley, J.; Walker, L. R.; Meek, J. Y.; Johnston, M.; Stellwagen, L.; et al. 2018. Marijuana use during pregnancy and breastfeeding: implications for neonatal and childhood outcomes. *Pediatrics*, 142(3).

Sayyadiharikandeh, M.; Varol, O.; Yang, K.-C.; Flammini, A.; and Menczer, F. 2020. Detection of novel social bots by ensembles of specialized classifiers. In *Proceedings of the 29th ACM international conference on information & knowledge management*, 2725–2732.

Sinnenberg, L.; Buttenheim, A. M.; Padrez, K.; Mancheno, C.; Ungar, L.; and Merchant, R. M. 2017. Twitter as a tool for health research: a systematic review. *American journal of public health*, 107(1): e1–e8.

Statistics Canada. 2024. Table 17-10-0009-01 Population estimates, quarterly. Statistics Canada Database. https://doi.org/10.25318/1710000901-eng.

The Canadian Institute for Health Information (CIHI). 2025. Induced Abortions in Canada. https://www.cihi.ca/en/induced-abortions-in-canada. Accessed January 13, 2025.

Van Draanen, J.; Krishna, T.; Tsang, C.; and Liu, S. 2019. Keeping up with the times: how national public health and governmental organizations communicate about cannabis on Twitter. *Substance abuse treatment, prevention, and policy*, 14: 1–7.

Ethics Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes
   (e) Did you describe the limitations of your work? Yes
   (f) Did you discuss any potential negative societal impacts of your work? Yes, see conclusions.
   (g) Did you discuss any potential misuse of your work? Yes, see conclusions and ethical statement.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? Not applicable
   (b) Have you provided justifications for all theoretical results? Not applicable
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? Not applicable
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Not applicable
   (e) Did you address potential biases or limitations in your theoretical framework? Not applicable
   (f) Have you related your theoretical results to the existing literature in social science? Not applicable
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Not applicable
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? Not applicable
   (b) Did you include complete proofs of all theoretical results? Not applicable
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Not applicable
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Not applicable
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes
   (b) Did you mention the license of the assets? Not applicable
   (c) Did you include any new assets in the supplemental material or as a URL? Yes
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? Yes
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? Yes
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? Yes
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Not applicable
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Not applicable
   (d) Did you discuss how data is stored, shared, and deidentified? Yes

Ethical Statement

This study adheres to established ethical standards for research involving publicly available data. The CHEO Research Ethics Board reviewed our project and granted an exemption in May 2021, recognizing that the study only collects and synthesizes publicly available data. In December 2021, we obtained approval to use the Twitter API for Academic Research.

To ensure transparency and accountability, we published a detailed research protocol (Cresswell et al. 2022) outlining our data collection process and research goals. For reproducibility, we share the collected data (Espín-Noboa et al. 2025), including metadata generated through our analysis (such as SBERT and Botometer scores and inferred tweet and user locations). To preserve anonymity and comply with Twitter's Terms of Service, we only provide CanPreg-related keywords per tweet, geo-location data limited to administrative level 2 (state in the USA and province in Canada), and the generated metadata. These measures ensure that our study remains reproducible while respecting data privacy and ethical research guidelines.

While using public data has significant advantages, analysis must proceed cautiously given the sensitive nature of the topics studied—in our case, CanPreg. To mitigate potential harms, our research employs aggregated data at the province and state levels, effectively preventing individual identification and reducing the risk of reinforcing stigma. The datasets we share are fully anonymized. Additionally, given that medical or recreational cannabis use is legal in all studied Canadian regions and most U.S. states, sharing insights into its online popularity and correlations with drug use or abortion is more beneficial than harmful. Furthermore, all socioeconomic indicators employed (such as drug use, poverty, and abortion rates) are derived from publicly available national statistics, posing no legal or reputational risks to individuals or communities. We acknowledge the potential misuse of such methodologies (e.g., for commercial purposes); thus, we emphasize the collective responsibility of the scientific community, policymakers, and governments to educate the public and continue rigorous investigation into CanPreg's effects on parents and newborns.

Appendix

Annotations and Ground-Truth

We created a "ground truth" dataset comprising 1,000 tweets and 200 users to determine thresholds for SBERT (Reimers and Gurevych 2019b) (CanPreg-related tweets) and Botometer (Sayyadiharikandeh et al. 2020) (non-bot accounts). Seventeen participants annotated at least one file, each containing 100 tweets and 20 users. The annotation process involved two independent annotators per tweet and user. In cases of disagreement, a second pair of annotators independently reviewed the items. Remaining disagreements were resolved in a final round, where four annotators jointly reached a consensus. Inter-annotator agreement for each round is reported in Table A1. Figures A3 and A4 provide examples of the Excel forms used for annotation, with each task organized by sheet.

We automated the removal of unrelated tweets using SBERT (Reimers and Gurevych 2019b) semantic search. Specifically, we defined five queries: q_1 = "cannabis during pregnancy," q_2 = "smoking weed while pregnant," q_3 = "the effects of cannabis on pregnant women," q_4 = "smoking or consuming drugs during pregnancy," and q_5 = "the effects of cannabis on newborns." SBERT calculated a semantic similarity score between each tweet and these queries. Ground-truth labels were then used to identify the minimum similarity score required for optimal performance. Figure A1 illustrates the precision-recall curves for each query, from which we selected thresholds that balanced precision and recall (i.e., the intersection point of the curves).

Botometer (Sayyadiharikandeh et al. 2020) was used to identify bot accounts. Using the ground-truth labels, we determined the optimal CAP threshold, as shown in Figure A2.

We applied both algorithms to the full tweet corpus, retaining tweets that met at least one of the five query thresholds ($q_1 \geq 0.43$, $q_2 \geq 0.48$, $q_3 \geq 0.42$, $q_4 \geq 0.42$, $q_5 \geq 0.43$) and users with a CAP score of ≤ 0.8.

		Krippendorff's α
Round 1	1000 tweets	0.56
	200 users	0.65
Round 2	189 tweets	0.49
	75 users	0.46
Round 3	46 tweets	1.00
	27 users	1.00

Table A1: **Inter-annotator agreement.** Krippendorff's α scores across three annotation rounds for tweet CanPreg relevance and bot detection tasks. Perfect agreement was achieved in the final consensus round.

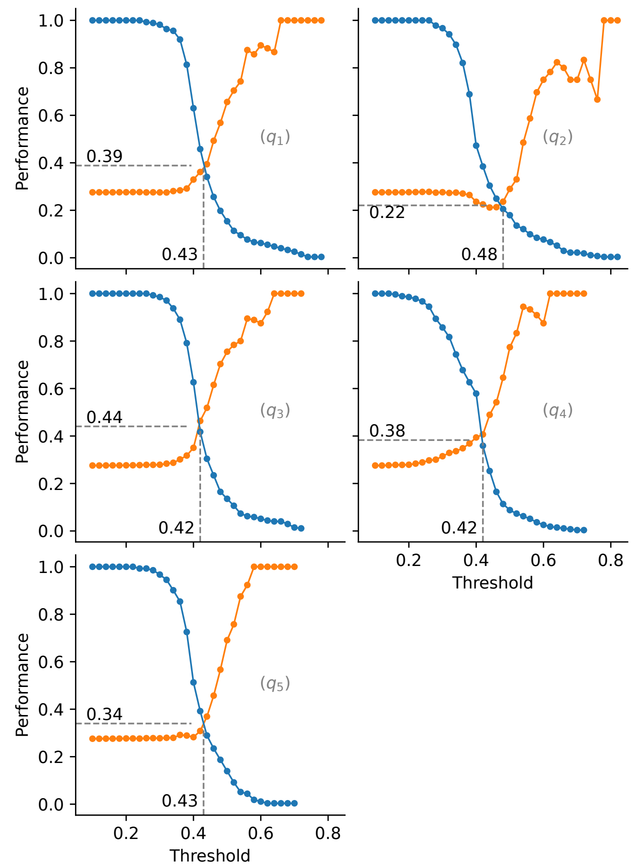


Figure A1: **SBERT threshold.** SBERT was employed to filter out non-CanPreg-related tweets. Using semantic search, it calculated similarity scores between each tweet and five predefined queries q_i. These scores (x-axis) were combined with ground-truth labels to determine the optimal threshold, identified at the point where precision and recall curves intersect, achieving balanced performance.

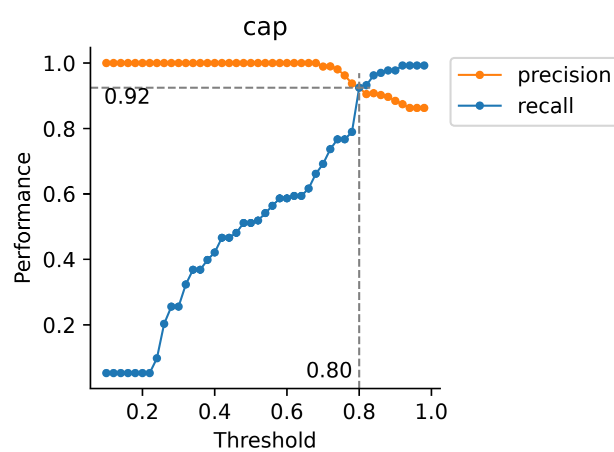


Figure A2: **Botometer threshold.** Botometer was used to estimate the likelihood of each of the 200 sampled users being a bot using a CAP score (x-axis). This precision-recall curve was used to select an optimal threshold of 0.8, achieving balanced precision and recall of 0.92.

```
ANNOTATING TWEETS

Dear participant,
This file contains 100 different tweets. Your task is to read each of them and
judge whether or not they are related to any of these topics:
- Cannabis use during pregnancy
- Smoking weed while pregnant
- The effects of cannabis on pregnant women
- Smoking or consuming drugs during pregnancy
- The effects of cannabis on newborns
- Kids, child, and youth smoking cannabis
- Medical cannabis for people
- Legalization of cannabis

If you think the tweet is related to any of the aforementioned topics, please
select or type "yes" (lowercase) under the "Is related?" column.

Have fun and thanks for volunteering!

DISCLAIMER: Some tweets might show explicit content or language.

Examples:
```

#	Text	Is related?
1	new book by dr. nice: recreational drugs and drugs used to treat addicted mothers. impact on pregnancy and breastfeeding. - dr. frank nice #marijuana #breastfeeding #lactation #ibclc #iamgold #goldlactation2018	yes
2	study: pregnancy and marijuana shouldn't mix url via @pioneerpress	yes
3	??? rt @ap: colorado lawmakers ponder whether to require warnings in pot shops about marijuana use by pregnant women: url	yes
4	this australian family was raided for cannabis oil (under doctor supervision) that saved their baby's life. how... url	yes
5	she like i left my weed at home . baby that is not a excuse	no
6	@iamcocov dope interview with @murdahbaby url	no
7	i'm not going to lie.. my baby mama's house is looking dope as hell right now lol	no
8	@permaconfused apparently they were blowing weed smoke at a pregnant woman so i think this is just an arsehole thing	no

```
Annotations:
```

#	Text	Is related?
1	icymi (how did we?!): 3 ways baby boomers will bolster the cannabis industry @mmjinvestor url @cassandradowell	-
2	dad kills baby in horror crash with cocaine and cannabis in blood url	-
3	gma said the weed helps me baby	-
...	...	...
98	from disney to smoking weed with wiz: life story of miley cyrus. my baby.	-
99	damn paul wall and baby bash got caught up this is why weed should be legalized in the united states as alcohol and outlaw cigarettes	-
100	how dumb could you be to fail a drug test for weed? unless it's like literally the fbi testing you baby, wyd.	-

Figure A3: **Annotation form for tweets.** A random sample of $1,000$ tweets was distributed across 10 Excel files, with each file assigned to two independent researchers for annotation as "Yes" (CanPreg-related) or "No" (not related). For tweets lacking consensus, additional annotators were engaged until agreement was reached, ensuring all tweets were consistently classified.

```
ANNOTATING USERS

Dear participant,
This file contains 20 different Twitter user accounts. Your task is to go to
their profiles (click on URL, see the picture, read the bio and a bunch of
tweets from their timelines) and judge whether they are humans  (i.e., someone
who writes his/her own tweets, retweets and likes other tweets,  and follows
and is followed by a reasonable number of people),  organizations (i.e., an
account that belongs to a brand, company, institute, and mostly spreads news),
or bots (i.e., accounts who mostly retweet, and their tweets are rarely personal).

If you think the user is a human (organization/bot), please select or type "yes"
under the "human?" (org?/bot?) column. If you are unsure, please select/type "maybe".
If the account does not exist, or is protected or it has been deactivated, please
choose "maybe" in all three categories. Note that the categories are non-exclusive,
meaning that a user can be a person and an organization at the same time.

Have fun and thanks for volunteering!

DISCLAIMER: Some accounts might show explicit content or language.

Examples:
```

#	Name	Username	URL	human?	org?	sbot?
1	*display_name*	*username*	https://twitter.com/username	yes	no	no
2	*display_name*	*username*	https://twitter.com/username	yes	yes	no
3	*display_name*	*username*	https://twitter.com/username	no	yes	yes
4	*display_name*	*username*	https://twitter.com/username	maybe	no	no
5	*display_name*	*username*	https://twitter.com/username	yes	no	no
6	*display_name*	*username*	https://twitter.com/username	maybe	no	maybe
7	*display_name*	*username*	https://twitter.com/username	yes	no	no
8	*display_name*	*username*	https://twitter.com/username	yes	no	no

```
Annotations:
```

#	Name	Username	URL	human?	org?	bot?
1	*display_name*	*username*	https://twitter.com/username	-	-	-
2	*display_name*	*username*	https://twitter.com/username	-	-	-
3	*display_name*	*username*	https://twitter.com/username	-	-	-
...	...	...	...	...	...	...
18	*display_name*	*username*	https://twitter.com/username	-	-	-
19	*display_name*	*username*	https://twitter.com/username	-	-	-
20	*display_name*	*username*	https://twitter.com/username	-	-	-

Figure A4: **Annotation form for users.** A random sample of 200 users was divided into 10 Excel files, each assigned to two independent researchers for classification as “human” (real person), “org” (organization or company), or “bot” (automated account). Disagreements were resolved by additional annotators to ensure consistent classification. Display and user names shown in this figure have been anonymized to protect privacy.

Additional Information for RQ1 (Geolocation)

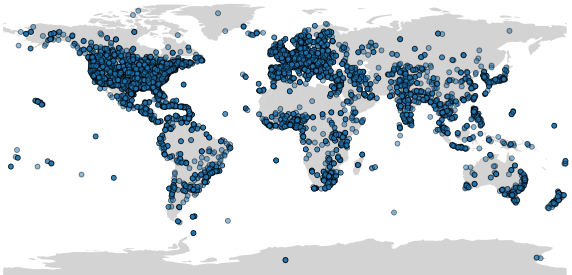
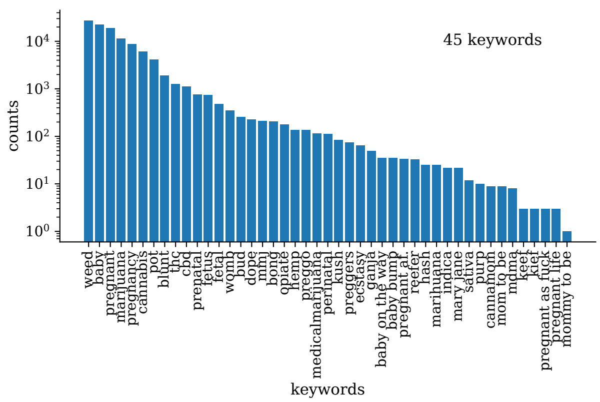
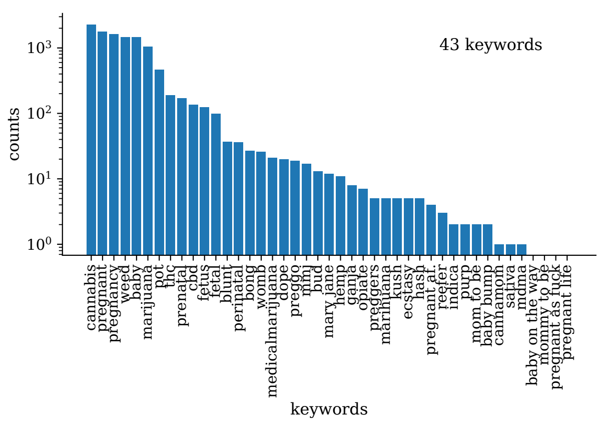


Figure A5: **Users posting CanPreg-related tweets.** All 46,366 geo-located valid users in our dataset. Most users are from USA (65%) followed by the UK (6, 8%), Canada (5, 4%), Australia (1, 3%), and France (1, 2%).

Additional Information for RQ2 (Topics)



(a) USA



(b) Canada

Figure A6: **Keyword frequency.** Keywords on the x-axis were used in our Twitter search to identify potential CanPreg-related tweets. The bars on the y-axis represent the number of distinct tweets containing each keyword.

(a) USA

(b) Canada

Figure A7: **Word counts.** Most common words in valid geo-located tweets from (a) the USA and (b) Canada. Sharing URLs, references to women, and smoking are prevalent in both countries. In contrast, terms related to breastfeeding and research appear relatively more frequently in Canada compared to the USA. Keywords used for Twitter search are excluded from this plot.

Additional Information for RQ3 (Surveys)

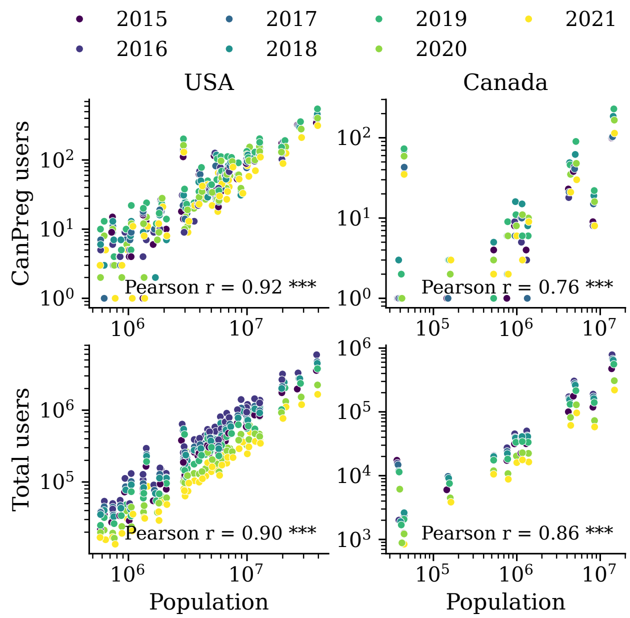
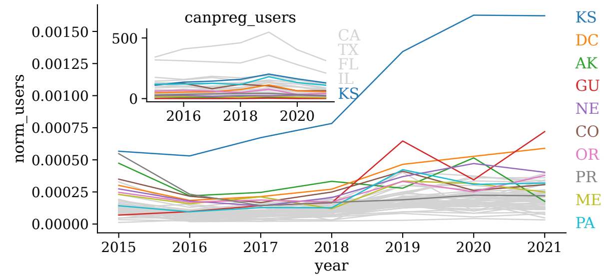
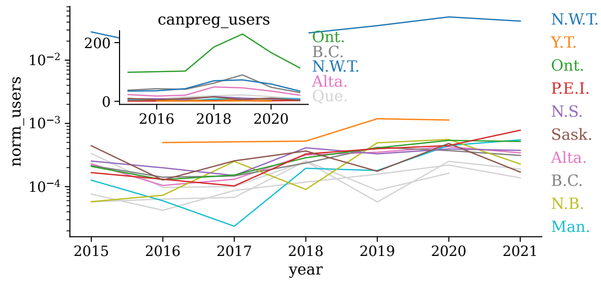


Figure A8: **Correlation between Twitter and population data.** Each column represents a country. The top panel shows the yearly count of unique users who posted at least one CanPreg-related tweet, while the bottom panel displays the total number of users tweeting in the respective year (colors). A positive trend is observed, indicating that larger regions tend to have more users overall and more CanPreg-related users. Pearson correlations were calculated across all regions and years in each panel, with all correlations being statistically significant ($p \leq 0.001$).

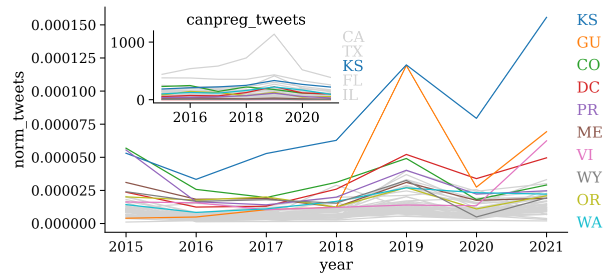


(a) USA

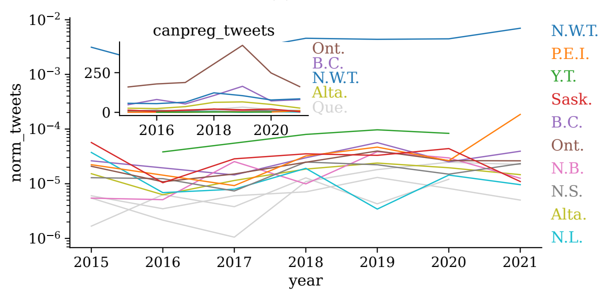


(b) Canada

Figure A9: **Number of users per year.** Normalized user counts per year and region are shown for (a) the USA and (b) Canada. Normalization is based on CanPreg raw counts (inset) divided by the total number of users posted each year. The top 10 regions (and top 5 in the inset) are highlighted. Regions are ranked in descending order by their total normalized user counts (raw counts in the inset).

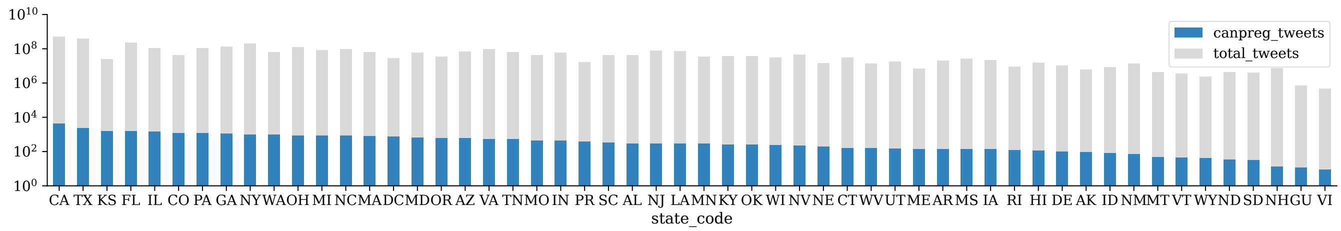


(a) USA

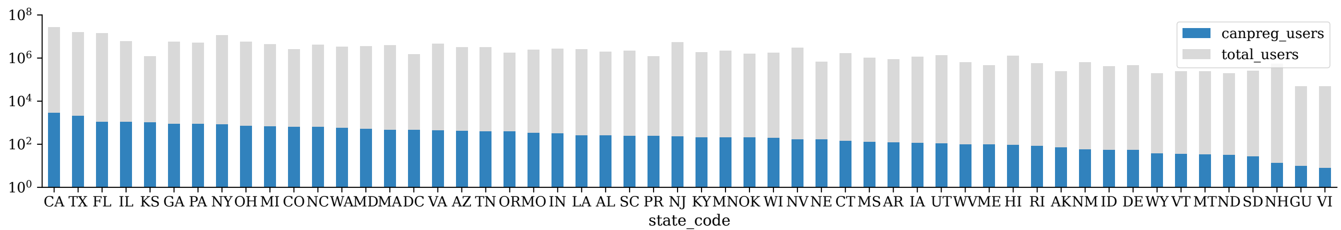


(b) Canada

Figure A10: **Number of tweets per year.** Normalized tweet counts per year and region are shown for (a) the USA and (b) Canada. Normalization is based on CanPreg raw counts (inset) divided by the total number of tweets posted each year. The top 10 regions (and top 5 in the inset) are highlighted. Regions are ranked in descending order by their total normalized tweet counts (raw counts in the inset).

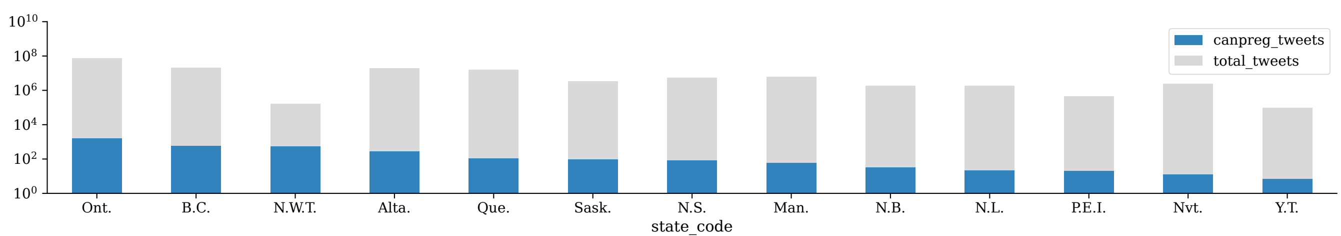


(a) Tweets

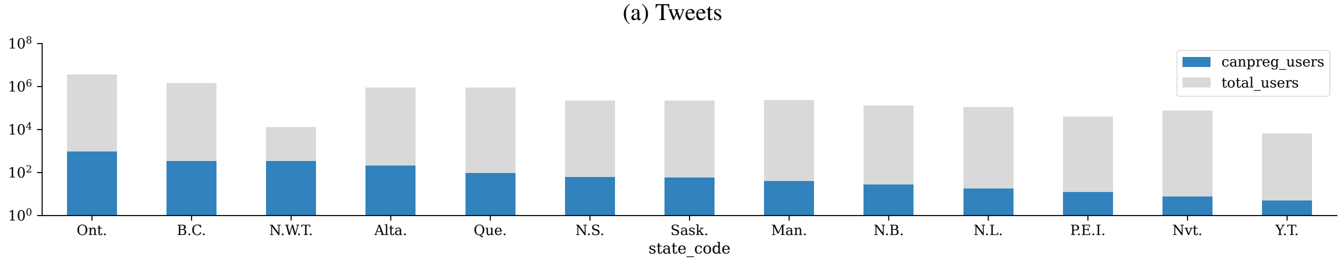


(b) Users

Figure A11: **Twitter data counts per state in USA.** (a) Number of CanPreg-related tweets compared to total tweets, and (b) Number of users discussing CanPreg compared to total users. Both plots highlight the distribution of activity across states, presented on a logarithmic scale.

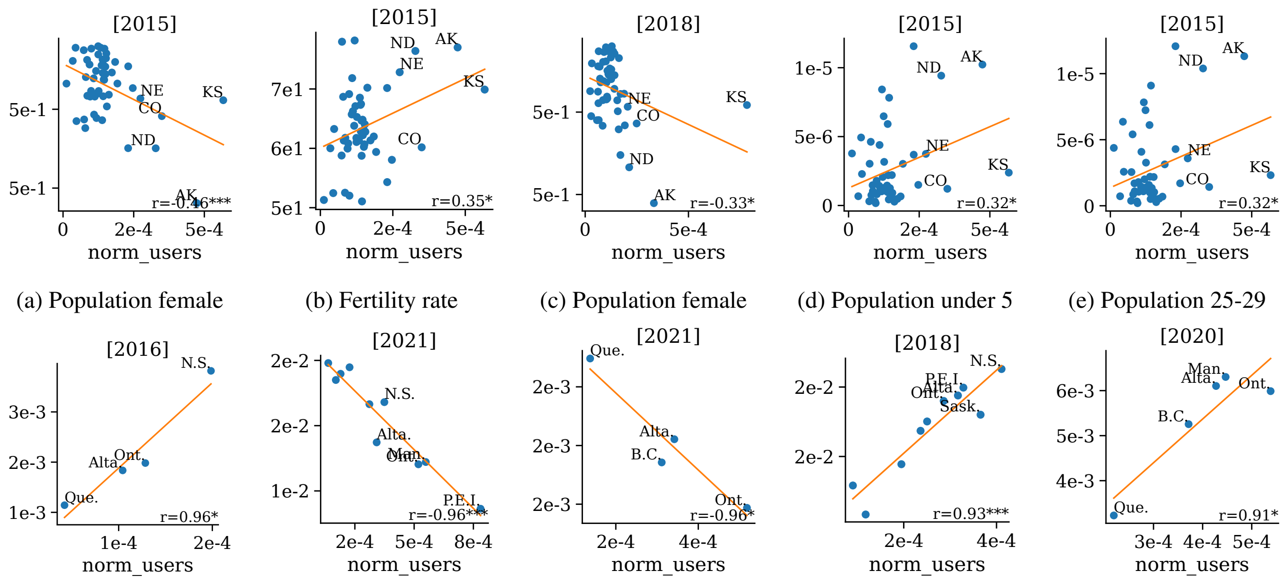


(a) Tweets



(b) Users

Figure A12: **Twitter data counts per province in Canada.** (a) Number of CanPreg-related tweets compared to total tweets, and (b) Number of users discussing CanPreg compared to total users. Both plots highlight the distribution of activity across Canadian provinces, presented on a logarithmic scale.



(a) Population female (b) Fertility rate (c) Population female (d) Population under 5 (e) Population 25-29

(f) Mental health F12-17 (g) Alcohol abuse F35-49 (h) Alcohol abuse 12-17 (i) Mental health 18-34 (j) Mental health 12-17

Figure A13: **Top-5 correlations between survey data and normalized CanPreg user counts.** Statistically significant correlations ($p \leq 0.05$) between normalized CanPreg user counts (x-axis) and various demographic and health metrics (y-axis). These include population characteristics, mental health, fertility rates, and alcohol abuse, analyzed across regions in the USA (a–e) and Canada (f–j) for the years 2015 and 2021.

Year	Metric (survey)	ρ	p
2018	*population_females	-0.33	0.02
	*population_males	0.33	0.02
	*years_45_to_49	-0.31	0.03
	*years_20_to_24	0.30	0.03
	*drug_deaths	-0.28	0.05
2016	*drug_deaths	-0.29	0.04
2015	*population_females	-0.46	0.00
	*population_males	0.46	0.00
	fertility_rate	0.35	0.01
	*years_under_5	0.32	0.02
	*years_25_to_29	0.32	0.03
	births_rate	0.31	0.03
	*births_total	0.30	0.03
	*years_20_to_24	0.30	0.03
	*years_5_to_9	0.30	0.04
	*years_30_to_34	0.30	0.04
	*years_10_to_14	0.28	0.05
	*drug_deaths	-0.28	0.05

Table A2: **Significant correlations between survey and CanPreg user data in the USA by year.** Metrics marked with (*) represent normalized counts adjusted for population size per region. Statistical significance was determined using a p-value threshold of $p \leq 0.05$. Results are presented in order of year and ranked by Pearson correlation strength.

Year	Metric (survey)	ρ	p
2020	births_avg_weight	0.29	0.04
2019	*insurance_private	0.30	0.04
2018	*poverty_total	-0.32	0.02
2015	*population_females	-0.29	0.04
	*population_males	0.29	0.04

Table A3: **Significant correlations between survey and CanPreg tweet data in the USA by year.** Metrics marked with (*) represent normalized counts adjusted for population size per region. Statistical significance was determined using a p-value threshold of $p \leq 0.05$. Results are presented in order of year and ranked by Pearson correlation strength.

Year	Metric (survey)	ρ	p
2021	*heavy_drinking_pop_females_35_to_49_years	-0.96	0.00
	*heavy_drinking_pop_12_to_17_years	-0.96	0.04
	*belonging_18_to_34_years	0.84	0.00
	cannabis_daily_females_65_years_and_over_percent	0.84	0.01
	*belonging_males_18_to_34_years	0.81	0.00
	*belonging_females_18_to_34_years	0.80	0.01
	*cannabis_daily_35_to_49_years	0.77	0.01
	cannabis_daily_35_to_49_years_percent	0.76	0.01
	*cannabis_daily_females_35_to_49_years	0.73	0.02
	cannabis_daily_females_35_to_49_years_percent	0.71	0.02
	*smoker_female_35_to_49_years	-0.71	0.02
	*heavy_drinking_pop_35_to_49_years	-0.69	0.03
	cannabis_use_65_years_and_over_percent	0.69	0.03
2020	*poor_mental_health_12_to_17_years	0.91	0.03
	cannabis_daily_50_to_64_years_percent	0.80	0.03
	*heavy_drinking_pop_females_35_to_49_years	-0.80	0.01
	*well_mental_health_females	-0.79	0.01
	*low_income_females_under_18_years	0.74	0.03
	*well_mental_health	-0.74	0.01
	*regular_healthcare_12_to_17_years	0.73	0.02
	*well_mental_health_females_50_to_64_years	-0.69	0.03
	*regular_healthcare_males_12_to_17_years	0.68	0.03
	*regular_healthcare_females_12_to_17_years	0.65	0.04
2019	*poor_mental_health_females_50_to_64_years	0.75	0.02
	*cannabis_daily_females	0.70	0.02
	cannabis_daily_females_percent	0.70	0.03
	*smoker_male_50_to_64_years	-0.66	0.04
	*cannabis_daily_35_to_49_years	0.64	0.05
2018	*poor_mental_health_18_to_34_years	0.93	0.00
	*poor_mental_health_females_35_to_49_years	0.85	0.01
	*poor_mental_health_females_18_to_34_years	0.75	0.02
	*poor_mental_health	0.65	0.04
2016	*poor_mental_health_females_12_to_17_years	0.96	0.04
	*poor_mental_health_males_50_to_64_years	0.79	0.02
	*poor_mental_health_males_35_to_49_years	0.73	0.04
	*poor_mental_health	0.73	0.03
	*poor_mental_health_males	0.69	0.04
	*poor_mental_health_females	0.69	0.04
	*smoker_female_18_to_34_years	-0.67	0.05
	*low_income_18_to_64_years	0.67	0.05
2015	*heavy_drinking_pop_males_18_to_34_years	0.69	0.03

Table A4: **Significant correlations between survey and CanPreg user data in Canada by year.** Metrics marked with (*) represent normalized counts adjusted for population size per region. Statistical significance was determined using a p-value threshold of $p \leq 0.05$. Results are presented in order of year and ranked by Pearson correlation strength.

Year	Metric (survey)	ρ	p
2021	*smoker_male_65_years_and_over	0.89	0.00
	*belonging_18_to_34_years	0.84	0.00
	*cannabis_daily_females_35_to_49_years	0.82	0.00
	*belonging_females_18_to_34_years	0.81	0.00
	*belonging_males_18_to_34_years	0.80	0.01
	cannabis_daily_females_35_to_49_years_percent	0.80	0.01
	*heavy_drinking_pop_females_18_to_34_years	0.78	0.01
	*smoker_male	0.75	0.01
	*smoker_65_years_and_over	0.75	0.01
	cannabis_daily_35_to_49_years_percent	0.73	0.02
	*heavy_drinking_pop_females_35_to_49_years	-0.71	0.02
	*cannabis_daily_35_to_49_years	0.70	0.02
	*cannabis_use_65_years_and_over	0.67	0.03
	*well_mental_health_35_to_49_years	-0.66	0.04
	cannabis_use_65_years_and_over_percent	0.65	0.04
2020	*poor_mental_health_males_18_to_34_years	0.96	0.04
	*poor_mental_health_females_12_to_17_years	0.89	0.04
	*well_mental_health	-0.79	0.01
	*well_mental_health_females	-0.72	0.02
	*well_mental_health_males_50_to_64_years	-0.68	0.03
	*well_mental_health_males	-0.68	0.03
	*well_mental_health_50_to_64_years	-0.67	0.04
	*low_income_under_18_years	0.65	0.04
	*regular_healthcare_12_to_17_years	0.64	0.05
2019	*smoker_male	-0.79	0.01
	*smoker_male_50_to_64_years	-0.77	0.01
	*smoker_35_to_49_years	-0.71	0.02
	*smoker_female_65_years_and_over	-0.70	0.02
	*poor_mental_health_females_50_to_64_years	0.68	0.04
	*smoker_male_35_to_49_years	-0.67	0.03
	*smoker_50_to_64_years	-0.65	0.04
2018	*poor_mental_health_18_to_34_years	0.75	0.01
2016	*poor_mental_health_35_to_49_years	0.92	0.00
	*poor_mental_health_females_65_years_and_over	0.84	0.02
	*smoker_female_18_to_34_years	-0.81	0.01
	*poor_mental_health_65_years_and_over	0.80	0.01
	*poor_mental_health_males_35_to_49_years	0.78	0.02
	*low_income_18_to_64_years	0.73	0.03
	*smoker_male_35_to_49_years	-0.70	0.04
	*low_income	0.67	0.05
2015	*poor_mental_health_females_35_to_49_years	0.68	0.04
	*heavy_drinking_pop_males_18_to_34_years	0.66	0.04

Table A5: **Significant correlations between survey and CanPreg tweet data in Canada by year.** Metrics marked with (*) represent normalized counts adjusted for population size per region. Statistical significance was determined using a p-value threshold of $p \leq 0.05$. Results are presented in order of year and ranked by Pearson correlation strength.

Table A6: **Semantic clusters.** Tweets were grouped into 80 semantically related clusters using a Sentence Transformer model. This table showcases 56 clusters based on tweet counts across both countries. We grouped 24 clusters as "Others" for babble or non-CanPreg tweets. Each cluster is manually assigned a general *topic* and a specific *sub-topic* for clarity. Examples are randomly selected and paraphrased tweets.

Cluster	# Tweets	Topic	Sub-topic	Example (paraphrased tweet)
1	6802	Research	CanPreg studies	Marijuana use doubles the risk of premature birth. [URL] via [MENTION]
5	316	Research	Baby boomers and cannabis use	Marijuana use is increasing among American baby boomers. [URL]
13	144	Research	Newborn cannabis testing	Baby soaps and shampoos can cause false positive marijuana tests. [URL]
28	64	Research	Opiate abuse	Preventing opiate-addicted babies during pregnancy. [URL] #pregnancydrugs
62	29	Research	Prenatal cannabis exposure	Prenatal cannabis linked to cognitive and behavioral issues. [URL] via [MENTION]
70	28	Research	Fetal harm from ecstasy use	Ecstasy in pregnancy may harm fetal motor skills. [URL]
3	532	Criticism	Criticism of CanPreg in general	I don't understand why pregnant women smoke weed or cigarettes smh.
11	147	Criticism	Criticism of specific CanPreg users	A girl at my school is pregnant and continues to smoke weed. That's just not okay.
20	88	Criticism	Prioritizing baby essentials	Your baby lacks essentials, but you have money for weed? That's irresponsible.
23	75	Criticism	Criticism of specific CanPreg users	[MENTION] You should know smoking weed during pregnancy is a big no.
27	67	Criticism	Criticism of parenting priorities	Posting weed photos with a 7-month-old is wrong.
31	58	Criticism	Criticism of CanPreg in general	Using substances while pregnant or parenting is irresponsible.
34	55	Criticism	Criticism of CanPreg in general	Smoking while pregnant and wondering why the baby cries.
39	43	Criticism	Smoking near infants	Can you stop smoking pot here? I have a baby in the apartment.
42	41	Criticism	Smoking near infants	I don't support smoking weed around a baby.
46	37	Criticism	Smoking near infants	Why smoke weed with a baby in the car? Seriously.
51	34	Criticism	Criticism of CanPreg in general	I can't support smoking cigarettes or weed during pregnancy. Just saying.
52	33	Criticism	Criticism of CanPreg in general	How are you pregnant and still smoking weed? #confused #smoking #pregnancy
61	30	Criticism	Smoking near infants	A mother blew marijuana smoke into her baby's mouth. [URL] What a world.
67	28	Criticism	Calling child protection services	You're pregnant and smoking weed? Hope CPS steps in
74	27	Criticism	Legal abortion vs. Ilegal marihuana	Why is smoking marijuana illegal, but ending a pregnancy isn't?
75	26	Criticism	Criticism of CanPreg in general	I support smoking weed, but doing it while pregnant is just not right, smh.
10	159	Experiences	Someone smoked while pregnant	My mom admitted to smoking weed while she was pregnant with me.
12	145	Experiences	Cannabis smell during pregnancy	When you're pregnant, the smell of weed is unbearable—just awful.
16	112	Experiences	Cannabis and sleep benefits	Weed helps me sleep so well, like a baby!
33	55	Experiences	Cannabis and sleep benefits	I wish I had some weed right now; it would help me sleep like a baby.
36	52	Experiences	CanPreg before prenatal checkups	My friend smoked weed today and has a pregnancy test tomorrow. Help?

Continues on next page

Cluster	# Tweets	Topic	Sub-topic	Example (paraphrased tweet)
43	41	Experiences	Cannabis and sleep benefits	That weed made me so sleepy—I slept like a baby.
45	37	Experiences	Morning sickness	Don't want to smoke pregnant, but it eases nausea. #life
47	35	Experiences	CanPreg is safe for the child	[MENTION] My friend smoked weed during pregnancy, and her son is fine.
48	35	Experiences	Choosing weed over pregnancy	If you can't quit weed for 9 months, it's a serious issue.
57	30	Experiences	CanPreg is safe for the child (sarcasm)	If you think weed makes babies smarter, reconsider having kids.
64	28	Experiences	Parent-child cannabis addiction	Smoking weed pregnant is like crack; baby can get addicted.
65	28	Experiences	Pregnancy and weed dreams	I keep dreaming I smoked weed and forgot I'm pregnant.
71	27	Experiences	Missing weed during pregnancy	Quit weed when pregnant, but miss it for stress relief. [URL]
9	173	News	Maternal neglect with cannabis	[MENTION]: Mom drives off with baby on car roof after smoking pot. [URL]
18	94	News	Infant cannabis overdose	Doctors attribute a baby's death to a marijuana overdose. [URL]
19	89	News	Maternal neglect with cannabis	Mother arrested and charged for giving marijuana to her baby. [URL]
29	59	News	Medical cannabis for infants	First baby treated with CBD oil at hospital. [URL]
38	44	News	Medical cannabis for infants	Medical marijuana: Can it stop this baby's seizures? [URL]
69	28	News	Child custody	Baby returned to marijuana-growing parents after custody seizure. [URL]
72	27	News	Cannabis warning labels for pregnancy	New bill: marijuana warnings for pregnant women. [URL]
77	26	News	Smuggling marihuana as baby bump	Woman caught with 'baby bump' hiding marijuana. [URL]
80	25	News	Criticism of CanPreg in general	Surgeon General warns teens, pregnant women: no marijuana. [URL]
7	213	Questions	CBD safety during pregnancy	Is using CBD to manage pregnancy symptoms safe? [URL]
17	104	Questions	Debates on CanPreg	Can you smoke weed while pregnant?
56	30	Questions	Debating CanPreg on Facebook	Saw girls on Facebook discussing weed while pregnant... wow.
76	26	Questions	CanPreg populaiton	Has anyone smoked weed during pregnancy?
22	78	Questions, Research	THC and fetal development	What about THC's effects on your baby? [URL]
49	35	Questions, Research, Experiences	Cannabis and sperm count	They say weed lowers sperm count, but I know daily smokers with many kids.
53	32	Questions, Research, Experiences	Fetal Alcohol Syndrome	[MENTION] Cannabis may not cause prenatal issues, unlike alcohol.
35	54	Criticism, Experiences	CanPreg and breastfeeding	I love breastfeeding, but I really miss weed—it's been almost a year since I quit.
59	30	Criticism, Experiences	Smoking near infants	Blew weed in a baby's face to calm them down!
63	28	Criticism, Experiences	Wine vs CanPreg	So many critique wine in pregnancy but smoke weed with tobacco.
44	39	Experiences, Research	CanPreg and breastfeeding	THC in breastmilk may cause drowsiness, but I didn't notice this with my baby.
54	31	Advertisements	Age-restricted product advertising	#cannabis 21+ only, avoid during pregnancy, don't drive, may impair judgment. [URL]

Continues on next page

Cluster	# Tweets	Topic	Sub-topic	Example (paraphrased tweet)
(Multi)	3612	Others	Lyrics, babble	Haven't smoked weed in 6 months. That first one will hit hard! [URL]
(None)	38819	Un-clustered	Various	Light up before intimacy—it's the trick to finishing without consequences. #magic