



Examining Cannabis, Tobacco, and Vaping Discourse on Reddit: An Exploratory Approach Using Natural Language Processing

Ryzen Benson^{1*}, Mengke Hu¹, Annie T. Chen², Shu-Hong Zhu³ and Mike Conway¹

¹ Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, United States, ² Department of Biomedical Informatics and Medical Education, University of Washington School of Medicine, Seattle, WA, United States, ³ Moores Cancer Center, University of California, San Diego, La Jolla, CA, United States

OPEN ACCESS

Edited by:

Waldemar W. Koczkodaj,
Laurentian University, Canada

Reviewed by:

Hsin-Yao Wang,
Linkou Chang Gung Memorial
Hospital, Taiwan

Agnieszka Szymańska,
Cardinal Stefan Wyszyński
University, Poland

Abeed Sarker,
University of Pennsylvania,
United States

*Correspondence:

Ryzen Benson
ryzen.benson@utah.edu

Specialty section:

This article was submitted to
Digital Public Health,
a section of the journal
Frontiers in Public Health

Received: 08 July 2021

Accepted: 26 November 2021

Published: 05 January 2022

Citation:

Benson R, Hu M, Chen AT, Zhu S-H
and Conway M (2022) Examining
Cannabis, Tobacco, and Vaping
Discourse on Reddit: An Exploratory
Approach Using Natural Language
Processing.
Front. Public Health 9:738513.
doi: 10.3389/fpubh.2021.738513

Background: Perceptions of tobacco, cannabis, and electronic nicotine delivery systems (ENDS) are continually evolving in the United States. Exploring these characteristics through user generated text sources may provide novel insights into product use behavior that are challenging to identify using survey-based methods. The objective of this study was to compare the topics frequently discussed among Reddit members in cannabis, tobacco, and ENDS-specific subreddits.

Methods: We collected 643,070 posts on the social media site Reddit between January 2013 and December 2018. We developed and validated an annotation scheme, achieving a high level of agreement among annotators. We then manually coded a subset of 2,630 posts for their content with relation to experiences and use of the three products of interest, and further developed word cloud representations of the words contained in these posts. Finally, we applied Latent Dirichlet Allocation (LDA) topic modeling to the 643,070 posts to identify emerging themes related to cannabis, tobacco, and ENDS products being discussed on Reddit.

Results: Our manual annotation process yielded 2,148 (81.6%) posts that contained a mention(s) of either cannabis, tobacco, or ENDS with 1,537 (71.5%) of these posts mentioning cannabis, 421 (19.5%) mentioning ENDS, and 264 (12.2%) mentioning tobacco. In cannabis-specific subreddits, personal experiences with cannabis, cannabis legislation, health effects of cannabis use, methods and forms of cannabis, and the cultivation of cannabis were commonly discussed topics. The discussion in tobacco-specific subreddits often focused on the discussion of brands and types of combustible tobacco, as well as smoking cessation experiences and advice. In ENDS-specific subreddits, topics often included ENDS accessories and parts, flavors and nicotine solutions, procurement of ENDS, and the use of ENDS for smoking cessation.

Conclusion: Our findings highlight the posting and participation patterns of Reddit members in cannabis, tobacco, and ENDS-specific subreddits and provide novel insights into aspects of personal use regarding these products. These findings complement epidemiologic study designs and highlight the potential of using specific subreddits to explore personal experiences with cannabis, ENDS, and tobacco products.

Keywords: marijuana, tobacco, smoking, electronic cigarettes, social media, natural language processing, Reddit, infodemiology

INTRODUCTION

Prior work has explored the broad facets of tobacco, electronic nicotine delivery systems (ENDS), and cannabis. This existing work has focused on various aspects of product use including the use, dual use (i.e., recent use of two product types), co-use (i.e., the simultaneous use of two product types), and user opinions of these different products (1–4). Furthermore, survey data has suggested that individuals frequently co-use these products, warranting the study of these products together rather than separately (5). Using consumer-generated data—in this context, textual data derived from social media services like Twitter and Reddit—continues to gain traction as a method to generate new insights into product use behavior. Using such approaches to study tobacco, ENDS, and cannabis could provide novel firsthand accounts on the use of these products and serve to complement existing survey-based approaches to study their use.

Tobacco Use in the United States

Since the release of the first US Surgeon General's report in 1964, the prevalence of combustible tobacco use has substantially declined in the US (6). Although smoking prevalence continues to decline, an estimated 16.7% of the US adult population still currently uses combustible tobacco products, and smoking remains the leading cause of preventable death in the US (6, 7). Combustible tobacco use is associated with a multitude of comorbidities, including but not limited to cardiovascular disease, chronic obstructive pulmonary disease (COPD), and numerous cancers (8). While these health detriments have increased the desire among smokers to quit, the 2015 National Health Interview Survey (NHIS) showed that only 7.4% of past year smoking cessation attempts were successful (9). Combustible tobacco use is more prevalent among males vs. females (20.1 and 13.6% respectively) and individuals between 25 and 44 years of age (7). Additionally, in the US smoking is more common among adults with lower educational attainment, lower-income status, and American Indian/Alaska Native individuals¹. Developing targeted tobacco cessation strategies among these individuals is a critical step in obtaining the goals outlined by Healthy People 2030 to reduce the prevalence of smoking in the U.S.² and, in turn, mortality due to smoking.

¹<https://cancercontrol.cancer.gov/brp/tcrb/monographs/monograph-22>

²<https://health.gov/healthypeople/objectives-and-data/browse-objectives/tobacco-use>

ENDS Use in the United States

ENDS or electronic cigarettes, are devices in which a nicotine solution that is often artificially flavored, is heated into a vapor that is then inhaled to simulate the act of smoking. ENDS devices include but are not limited to vape pens (pen-shaped nicotine vaporizers), mods (modifiable nicotine vaporizers), pod-mods (ENDS devices with disposable nicotine pods), and vaporizers (devices with refillable tanks for nicotine solutions)³. These devices differ mechanistically and in nicotine delivery, but were initially developed as a cessation aide that resembles the feel and experience of smoking cigarettes (10). Studies have found that the use of ENDS among adult smokers increased the rate of quit attempts and helped those who did attempt to quit achieve a higher rate of sustained abstinence (11–14). In recent years, ENDS use has continually increased, especially among adolescents (15). In 2019, it was estimated that ~4.5% of US adults used an ENDS device, compared to 10.5 and 27.5% of US middle school and high school students, respectively (7, 16). This increase in popularity is largely attributable to the emergence of pod mods, ENDS devices with sleek designs that mimic the nicotinic delivery of combustible cigarettes (17, 18). Although adults commonly use ENDS devices to aid in smoking cessation (6), studies have suggested that adolescents and young adults may be using ENDS devices recreationally rather than to aid smoking cessation (19, 20). Further, a recent study demonstrated that from 2014 to 2018, the adolescent age of initiation for ENDS devices continued to decrease (21), meaning that adolescents were beginning ENDS use at younger ages. These use patterns have sparked great concern about the potential for youth users to develop nicotine dependence, subsequent nicotine addiction, and a later transition to combustible cigarettes (22–24). In addition to the concerns of youth nicotine exposure, the long-term effects of ENDS use have yet to be ascertained, an outcome that can only be observed prospectively with future study.

Cannabis Use in the United States

Cannabis (marijuana) has long been stigmatized as an illicit drug in the US, despite the demonstrated health benefits associated with its use, including pain relief, chronic disease management, and stress reduction (25). However, there is also substantial evidence regarding the potential harms associated with cannabis use. Some of these detrimental health effects include neurological structure alterations, the onset of mental

³https://www.cdc.gov/tobacco/basic_information/e-cigarettes/pdfs/ecigarette-or-vaping-products-visual-dictionary-508.pdf

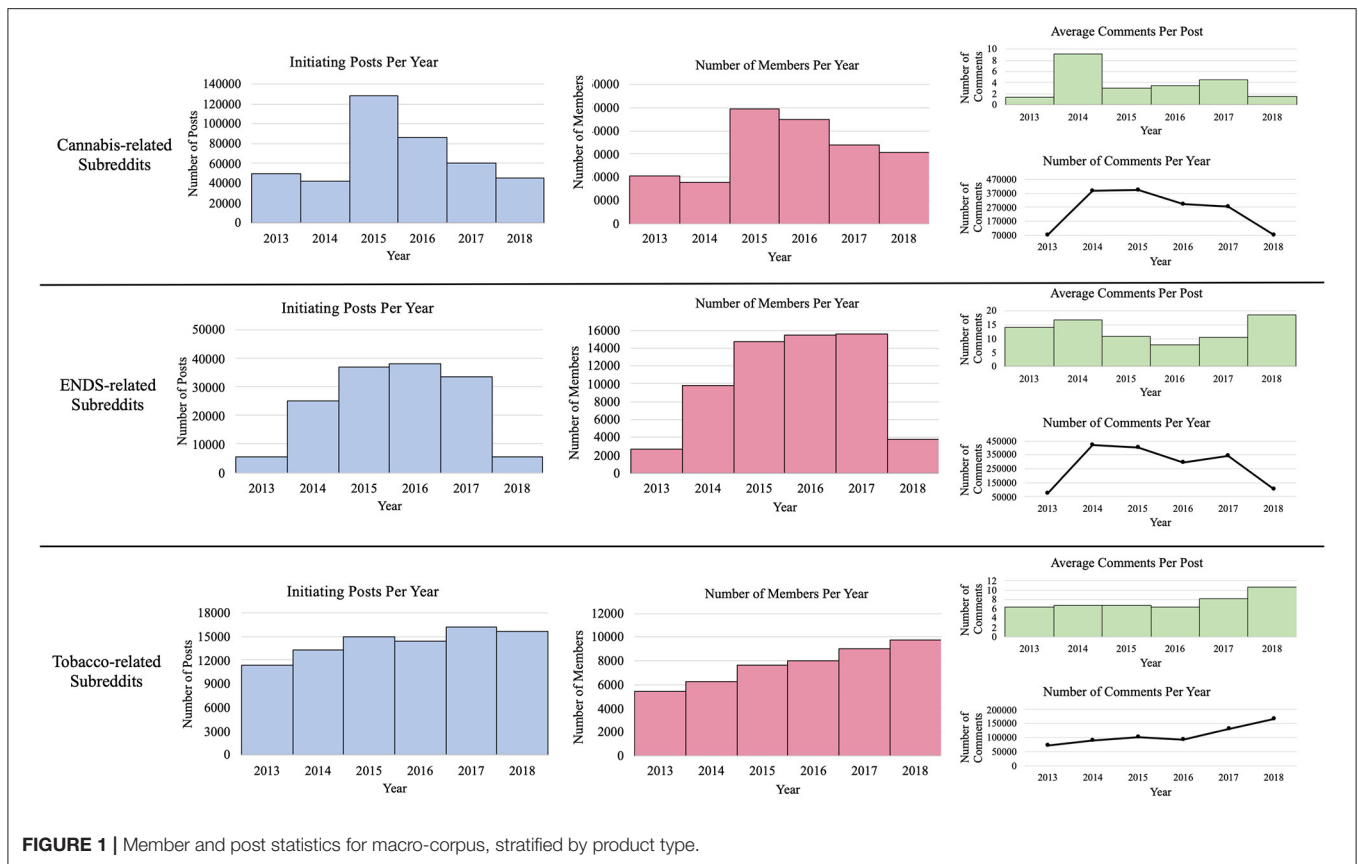


FIGURE 1 | Member and post statistics for macro-corpus, stratified by product type.

health disorders, cannabis dependency development, and the onset of various chronic conditions (26). From 2002 to 2014, significant increases in cannabis use and initiation were observed among US adults (27). Moreover, following the November 2020 US elections, 19 states and two US territories have now legalized recreational cannabis use and 35 states have legalized medical cannabis^{4,5}, where states with legalized cannabis tend to show a higher prevalence of use than states without legalized recreational cannabis (28, 29). In addition to the changing legislation surrounding medical and recreational cannabis, the forms and potency of cannabis available on the market continue to change as well. Tetrahydrocannabinol (THC) is the primary psychoactive cannabinoid found in cannabis and the determining compound of cannabis potency. Lab tests of cannabis seized by the Drug Enforcement Administration (DEA) showed that the average THC content had increased 4-fold from 3% in 1980 to 12% in 2012 (30). This high potency, accompanied by the various methods of ingestion (i.e., bud/flower, edibles, topicals, dab pens, extracts) (31), as well as the conflicting evidence regarding the health effects of its use, highlight the critical need for further study of user habits and perceptions of cannabis.

⁴<https://www.usnews.com/news/best-states/articles/where-is-marijuana-legal-a-guide-to-marijuana-legalization>

⁵https://www.medicinenet.com/how_many_states_legalized_medical_marijuana_2021/article.htm

Using Social Media in Public Health Research

For over a decade, the popularity of social media sites continues to rise, as do the functionalities of these sites (e.g., Instagram for posting pictures, Twitter for tweeting short posts, TikTok for posting short videos). Thanks to the development of social media-specific application programming interfaces (APIs), publicly available data posted on these sites can be collected for analysis by end users. As a result, public health researchers have leveraged this available data to study numerous aspects of public health. The research questions that can be explored are largely dependent on the specific site used to obtain data for analysis, as certain social media sites are better suited for specific research questions due to the structure of the website, community posting guidelines put in place, and limits posed by post structure.

Reddit⁶ is a popular social media site in which members under a self-chosen username post discussions within a subreddit (i.e., a forum centered around a common theme such as smoking cessation or cannabis use). As a result of this anonymity, members often discuss sensitive and oftentimes stigmatized health behaviors and conditions (32). Furthermore, unlike other social media sites such as Twitter and Facebook, Reddit posts are often more verbose, and since they are aggregated into subreddits centered around common themes of interest,

⁶<http://www.reddit.com>

the development of Reddit-specific APIs have resulted in an increasing body of literature leveraging Reddit to study aspects of substance use. Reddit has been used to study the effects of the COVID-19 pandemic on opioid use patterns and disruptions to treatment (33, 34), emerging forms of cannabis discussed in subreddits (35), potential adverse health effects associated with specific JUUL pod flavors (36), and to study linguistic patterns characteristic of alcohol and tobacco abstinence (37), to name a few. Further, we recently published results of natural language processing (NLP) pipelines developed to identify the prevalence of tobacco, cannabis, and ENDS mentions within tobacco, cannabis, and ENDS-related subreddits (38). However, that prior study was primarily concerned with the computational identification of tobacco, cannabis and ENDS mentions in Reddit posts, hence we did not further analyze these posts for the topics discussed within. Therefore, there remains need to investigate the topics of discussion frequent to tobacco, cannabis, and ENDS-centric subreddits.

To address this gap, our study uses qualitative and computational methods to explore member experiences and discussions in eight different subreddits related to cannabis, ENDS, and tobacco product use and to identify differentiating characteristics of these subreddits for assisting with the future study of these products on Reddit. We first explore member-specific experiences as well as general discussion of the three product types by manually annotating posts made in these eight subreddits. Second, we apply Latent Dirichlet Allocation (LDA) topic modeling to our macro-corpus (i.e., posts made in these subreddits over a 5-year span) to determine emerging themes in these posts through an unsupervised manner. Last, we create word clouds for each product (e.g., cannabis, tobacco, and ENDS) based on the vocabulary of our annotated corpus to compare the broad themes discussed in our annotated corpus vs. those topics identified by topic modeling. By understanding the discourse related to these three product types on Reddit, our work serves to identify the use patterns, behaviors, and common topics associated with these products, complementing existing methods for studying differences and similarities in product use behavior.

METHODS

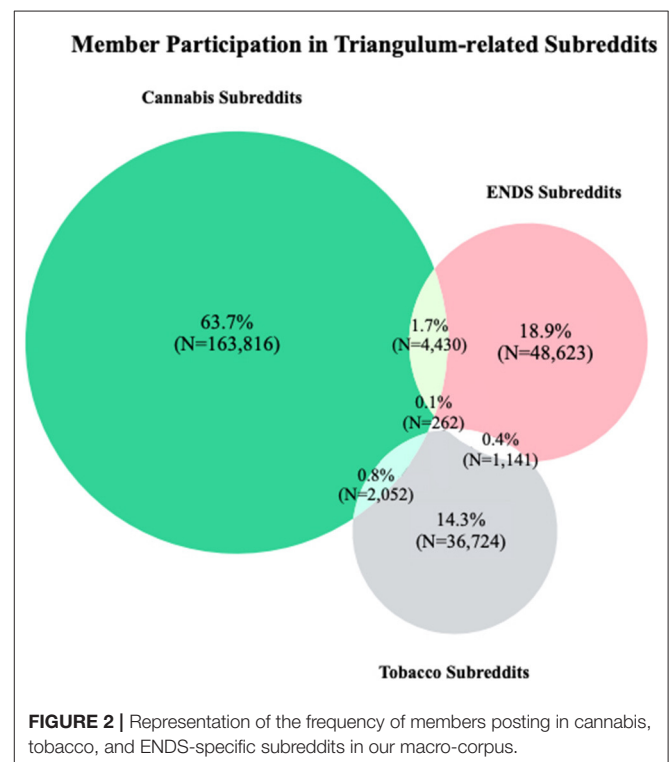
Data Collection

There are numerous subreddits focused on cannabis, ENDS, and tobacco, all of which serving a different purpose and with varying levels of member participation. Therefore, we selected a subset of subreddits with at least 50,000 members, demonstrating adequate member engagement within the subreddit, and providing a sufficient sample size to carry out our research objectives. Using the pushshift.io API (39, 40), we downloaded all available data from eight subreddits related to cannabis, ENDS, and tobacco between January 2013 to December 2018. These subreddits consisted of r/Vaping, r/electronic_cigarette, r/Vaping101, r/weed, r/trees, r/Marijuana, r/Cigarettes and r/stopsmoking. Prior work has observed frequent topic drift—a phenomenon where conversation in forums drift from the topic of interest to a different topic, particularly among online forums dedicated to health-related subjects (41). To mitigate topic

drift, and in an attempt to observe member-specific experiences and discussion, we only analyzed initiating posts in this study and did not evaluate subsequent comments. Our resulting macro-corpus consisted of 643,070 initiating posts. **Figure 1** presents a breakdown of the posts, subsequent comments, and members contained in our macro-corpus. We then stratified the macro-corpus by substance into cannabis-specific subreddits (i.e., r/Marijuana, r/weed, r/trees), tobacco-specific subreddits (i.e., r/stopsmoking, r/Cigarettes), and ENDS-specific subreddits (r/Vaping101, r/Vaping, r/electronic_cigarette). **Figure 2** provides the frequency of members participating in subreddits based on product type.

Annotation Scheme Development

Before analyzing our Reddit posts, we developed an annotation scheme to capture the various attributes and themes related to Reddit member experiences of ENDS, tobacco, and cannabis. To develop, refine, and evaluate the annotation scheme, we randomly selected a subset of 950 initiating posts from our macro-corpus. Based on prior work (42, 43), we developed an initial list of cannabis, ENDS, and tobacco-related keywords (e.g., tobacco, electronic cigarette, marijuana). As the brand names and terms used to describe tobacco products and cannabis are frequently evolving (35, 44), we used the Gensim Python library (45) to train a neural network-based algorithm, Word2Vec, on the entirety of the macro-corpus to identify additional keywords synonymous or plesionymous (i.e., close to synonymous) to our initial keywords. To ensure that we developed an annotation scheme capturing the attributes of interest (e.g., firsthand use of



combustible tobacco, historical cannabis use by someone else) for our study, we filtered posts that did not contain at least one of these keywords of interest. The resulting 280 cannabis, ENDS, and tobacco-related posts were then divided into seven batches for interrater agreement and annotation scheme development. Using the eHost annotation tool (46), authors MH, RB, AC, and MC annotated each batch of posts according to the corresponding version of the annotation scheme. Following each round of annotation, any discrepancies between annotators were discussed, resolved, and an adjudicated annotation set was created. After the seven annotation rounds, the interrater agreement reached an F-score of 0.83, a measure used as a surrogate for Cohen's Kappa and indicating a strong level of agreement among annotators (47–49). Our final annotation scheme can be found in the **Supplementary Material**.

Manual Annotation

Having developed and evaluated our annotation scheme, we identified a subset of posts for manual annotation. Due to the observation that some members created initiating posts more frequently than others, and to ensure diversity in the posting patterns represented, we stratified our members into five bins according to the frequency of their initiating posts and selected a random sample of members from each bin. As seen in **Figure 3**, 80 members created between 4 and 10 initiating posts, 30 members created between 11 and 50 initiating posts, eight members created between 51 and 100 initiating posts, four members created between 100 and 218 initiating posts, and two members created between 218 and 1,000 initiating posts. Altogether, our sample comprised of 124 Reddit members and their 2,630 initiating posts from the macro-corpus. Author RB annotated the majority of these posts according to the annotation scheme, while author MC annotated a small subset of the posts to ensure the quality of the agreement was maintained. Posts may have contained one or more attributes from our annotation scheme; therefore, annotations are not necessarily mutually

exclusive. The resulting proportions and frequencies from our manual annotation can be observed in **Table 1**.

Data Pre-processing

To pre-process our data for computational analyses, we first converted all of the Reddit posts to lower case. Using the Natural Language Toolkit (NLTK) (50), a widely used Python module for text analysis, we then split (i.e., tokenized) each post into individual word tokens. Lastly, we iterated through each token and removed any stop words. Stop words (e.g., “the,” “is,” “what”) are words that are common (51) but may not necessarily contribute semantic meaning to a corpus.

Latent Dirichlet Allocation and Word Clouds

While our manual annotation provides a comprehensive view of member experiences with cannabis, ENDS, and tobacco on Reddit, manually annotating our entire macro-corpus ($N = 643,070$) is not feasible. Consequently, topic modeling proves to be an invaluable method for studying commonly discussed topics within a large textual corpus. Latent Dirichlet Allocation (LDA) is one of these unsupervised machine learning techniques used to identify topics discussed in a corpus. Topics in an LDA model are represented by the grouping of words that are similar or co-occur throughout the corpus (51–53). LDA has been used to study several public health topics, including but not limited to infectious diseases, obesity, vaccinations, and health communications (32, 54–56), and has been shown effective in identifying salient topics within unstructured text sources such as social media. To develop our LDA models, we again used the Gensim library—as it supports various topic modeling techniques (45) and pyLDAvis to assist with the interpretation of the most salient terms for each topic (57). Then using our stratified macro-corpus (i.e., cannabis-specific subreddits, ENDS-specific subreddits, and tobacco-specific subreddits), we developed three distinct LDA models, one for each product type.

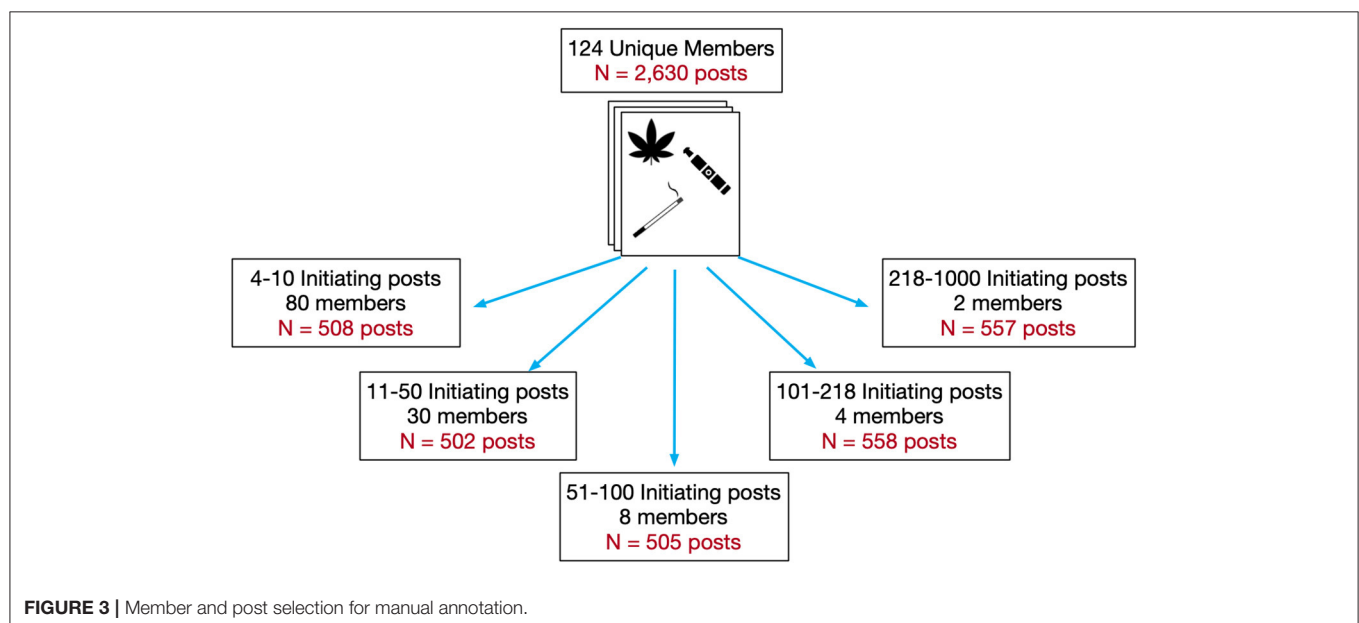


FIGURE 3 | Member and post selection for manual annotation.

TABLE 1 | Frequencies and percentages of posts from the manual annotation.

	r/trees N = 658	r/weed N = 2	r/Marijuana N = 875	r/stopsmoking N = 166	r/Cigarettes N = 37	r/Vaping N = 343	r/electronic_cigarette N = 67
Experiencer							
First person experience	285 (43.3)	1 (0.5)	10 (1.1)	122 (73.4)	16 (43.2)	205 (59.7)	42 (62.6)
Experience other	20 (3.0)	0 (0.0)	0 (0.0)	12 (7.2)	0 (0.0)	7 (2.0)	1 (1.4)
General discussion	486 (73.8)	2 (100.0)	872 (99.6)	130 (78.3)	32 (86.4)	225 (65.5)	63 (94.0)
Temporality							
Historical	14 (2.1)	0 (0.0)	1 (0.1)	46 (27.7)	1 (2.7)	3 (0.8)	4 (5.9)
Present	286 (43.4)	1 (0.5)	10 (1.1)	102 (61.4)	15 (40.5)	202 (58.8)	42 (62.6)
Future	6 (0.9)	0 (0.0)	1 (0.1)	2 (1.2)	0 (0.0)	17 (4.9)	1 (1.4)
Product							
Marijuana	655 (99.5)	2 (100.0)	874 (99.8)	2 (1.2)	1 (2.7)	3 (0.8)	0 (0.0)
Combustible tobacco	33 (5.0)	1 (0.5)	6 (0.6)	165 (99.3)	37 (100.0)	14 (4.0)	8 (11.9)
ENDS	5 (0.7)	0 (0.0)	0 (0.0)	8 (4.8)	0 (0.0)	341 (99.4)	67 (100.0)

Annotations are not necessarily mutually exclusive.

We manually varied LDA model hyperparameters (i.e., number of iterations and batch size to train the models) to determine the optimal number of topics (k) to identify themes frequent to each product type. We then manually inspected the resulting terms with these varying k values to observe coherence between and ultimately determine the relevant number of k topics, as well as the best label that encompasses the observed terms. Lastly, using our annotated corpus, we developed a word cloud for each product type using the WordCloud Python library (58). These analyses together, allowed us to compare commonly discussed topics in both our annotated corpus and the larger macro-corpus.

Ethical Considerations

This study was determined to be exempt from review by the University of Utah Institutional Review Board (IRB#00076188). To protect Reddit member privacy, we have refrained from including usernames in this paper. Further, all quotations used are synthesized from multiple examples.

RESULTS

Macro-Corpus

As seen in **Figure 1**, we saw a large increase in member participation and the number of posts made in cannabis-specific subreddits in 2015, followed by decreases through 2018. In ENDS-specific subreddits, we observed annual increases in the number of posts within these subreddits between 2013 and 2016, followed by a slight decrease in volume in 2017 and a drastic decrease in 2018, whereas tobacco-specific subreddits saw continual increases in posts made and member participation per year between 2013 and 2018. In total, the 643,070 posts in our macro-corpus were posted by 257,048 unique members across all eight subreddits of interest. Some members deactivated their Reddit accounts between the date of the post and when we collected the Reddit posts. Therefore, those usernames are not available and are referred to as “[Deleted]” by the API.

As shown in **Figure 2**, 63.7% of members ($N = 163,816$) posted exclusively in cannabis-specific subreddits, 18.8% of

members ($N = 48,623$) posted exclusively in ENDS-specific subreddits, and 14.3% of members ($N = 36,724$) posted exclusively in tobacco-specific subreddits. The remaining ~3% of members posted in multiple product-specific subreddits. We observed 4,430 members posting in both cannabis and ENDS-specific subreddits, 2,052 members posting in both cannabis and tobacco-specific subreddits, 1,141 members posting in both ENDS and tobacco-specific subreddits, and 262 members who posted in cannabis, tobacco, and ENDS-specific subreddits.

Manual Annotation

The complete results of our manual annotation can be seen in Table 1. Of the 2,630 posts that we manually annotated, 2,148 (81.6%) posts contained at least one mention of either cannabis, ENDS, or tobacco products. 1,537 (71.5%) of these posts mentioned cannabis, 421 (19.5%) posts mentioned ENDS, and 264 (12.2%) posts mentioned tobacco. Only three posts explicitly mentioned the dual use of cannabis, ENDS, and tobacco products, and 11 posts explicitly mentioned transitions between these products. Two of the three posts mentioning dual use were made by one member. In these posts, the member explicitly mentioned their personal experiences with these products and stated current use of all three products. Upon closer examination of the posts containing transitions between products, six posts were made in the r/Vaping subreddit, three were made in the r/stopsmoking subreddit, and two posts were made in the r/electronic_cigarette subreddit. All 11 posts were related to past, ongoing, or planned smoking cessation attempts or associated relapse events, and none of these posts mentioned cannabis. Four posts discussed relapse from ENDS devices to combustible cigarettes, while three posts documented ongoing smoking cessation attempts using ENDS devices. Additionally, one post discussed a failed past cessation attempt using ENDS, two posts discussed contemplating the use of ENDS to facilitate a future cessation attempt, and one post discussed a relapse from ENDS devices to cigarettes and then subsequently reinitiating ENDS use.

Within r/Marijuana, 16 members posted general discussions (e.g., news stories, questions, non-personal experiences with cannabis), and three members posted self-experiences with cannabis compared to r/trees where 70 members posted self-experiences with cannabis, and 65 members posted general discussions, suggesting that members are more likely to post personal experiences with cannabis in the r/trees subreddit compared to r/Marijuana. Of the 311 posts containing mentions of experiences with cannabis, 16 (0.5%) posts mentioned former use, 299 (96.1%) posts mentioned current use, and 7 (0.2%) posts mentioned potential use in the future. Of the 185 posts containing mentions of experiences with tobacco, 53 (28.6%) posts mentioned former use or past cessation attempts, 158 (85.4%) posts mentioned current use or current cessation attempts, and 7 (3.7%) posts mentioned potential use or contemplation of potential cessation attempts in the future. Of the 261 posts containing mentions of ENDS experiences, 13 (4.9%) posts mentioned former use, 252 (96.5%) posts mentioned current use, and 19 (7.2%) posts mentioned potential use in the future.

Latent Dirichlet Allocation and Word Cloud Analysis

In the LDA models applied to our macro-corpus, we observed six frequently discussed topics in cannabis-specific subreddits, four frequent topics in ENDS-specific subreddits, and two common topics in tobacco-specific subreddits. Cannabis-related topics included the legalization of medicinal and recreational cannabis, experiences with and recreational use of cannabis, the methods and forms of cannabis, health effects and uses of cannabis, and the cultivation of cannabis plants. ENDS-specific subreddits often held discussions of different flavors and nicotine solutions, accessories and parts, procurement of ENDS devices, and the use of ENDS for smoking cessation. And tobacco-specific subreddits often contained discussions of different brands and types of combustible tobacco and current, past, and planned smoking cessation attempts. **Figure 4** provides the common terms used to derive each topic label from our LDA analysis and synthetic textual examples of these topics.

The observed topics from the LDA analysis of our macro-corpus closely mimic the terms from the word clouds of our annotated posts, as seen in **Figure 5**. In the word clouds developed from our annotated corpus, the cannabis-specific word cloud consisted of terms including “cannabis,” “high,” “legalization,” and “medical.” In the tobacco-specific word cloud, we observed terms such as “smoke,” “quit,” and terms that may suggest temporality components of smoking cessation such as “time,” “today,” “month,” and “year.” Furthermore, in the ENDS-specific word cloud, we observed terms such as “flavor,” “tank,” “mod,” and “juice.”

DISCUSSION

Principal Findings

Reddit is a popular social media platform in which members post content in communities (subreddits) centered around everyday topics. Reddit posts are typically more verbose than other social media sites (e.g., Twitter, Instagram) and have numerous

subreddits dedicated to the discussion of various aspects related to cannabis, tobacco, and ENDS products. Leveraging this structure, we studied eight subreddits focused on cannabis, tobacco, and ENDS, identifying emerging themes and analyzing how these themes differ between product types.

In cannabis-specific subreddits, we observed a large increase in the number of posts and the number of Reddit members within these subreddits in 2015. This observation is likely a result of the increase in proposed cannabis legislation in the mid 2010's, a controversial topic of debate throughout the United States⁷. Our topic analysis of these subreddits also supports this finding as cannabis policy was one of the emerging themes discussed throughout our macro-corpus of posts. In addition to legislation and policy, members in cannabis-specific subreddits frequently talked of personal experiences with cannabis, the health effects of cannabis use, methods and forms of cannabis, and the cultivation of cannabis. Upon manual annotation and a closer examination of posts containing cannabis-related discussion, the subreddits r/weed and r/Marijuana appear to harbor more general discussion of cannabis, including discussion related to legalization and regulation, a finding also observed in prior work (32). The vast majority of posts containing personal experiences of cannabis were found in the subreddit r/trees. Within these experiential cannabis-related posts, members often talk about methods of consuming cannabis, experiences while under the influence, stories of cannabis use, and were typically centered around present use of cannabis. While previous studies have used other social media sites to explore the use of cannabis (59, 60), and a prior study compared cannabis use as discussed on one subreddit (35), no study, to the best of our knowledge, has studied aspects of cannabis use across multiple subreddits. Consequently, these findings are significant contributions—demonstrating the potential of Reddit data to explore opinions and use patterns of cannabis while also providing guidance as to which subreddits are best for studying personal use experiences versus general discussion. Based on our findings, future work that seeks to leverage Reddit for studying cannabis use and perceptions should do so using the r/trees subreddit.

From 2013 to 2018, we observed continual increases in the number of posts made and members participating in tobacco-specific subreddits. This sustained volume of posts within these subreddits may be a result of the observed increase in desire among smokers to quit in recent years (61), but may be an artifact of increasing usage of Reddit since 2012⁸. These patterns are consistent with our topic modeling analysis in which we observed smoking cessation frequently being discussed throughout posts made in tobacco-specific subreddits. In addition to aspects of cessation, we frequently observed discussions among current smokers regarding their favorite brands and types of combustible tobacco (e.g., Marlboro, unfiltered). These discussions were mostly housed in r/Cigarettes, a subreddit dedicated to discussions of cigarettes among smokers. Conversely, posts in r/stopsmoking were typically from current smokers attempting or contemplating a quit attempt, individuals

⁷<https://www.pewresearch.org/politics/2015/04/14/in-debate-over-legalizing-marijuana-disagreement-over-drugs-dangers/>

⁸<https://backlinko.com/reddit-users>

experiencing relapse from a smoking cessation attempt, or individuals who successfully quit smoking and sustained their abstinence. In this subreddit, members documented their quit journey, including withdrawal symptoms and side effects, while also asking for advice to assist in their quit attempts. These findings reinforce the work of Chen et al., in which the investigators evaluated thematic elements expressed in the stopsmoking subreddit (62).

From 2013 to 2017, the frequency of Reddit members posting in ENDS-specific subreddits increased. These observed increases in ENDS-related posts may be a result of the increasing popularity of ENDS devices in the mid 2010's but may also reflect the increase in Reddit traffic since 2012, as stated prior. Pew research surveys show that 22% of Reddit members are young

adults between the ages of 18 and 29⁹, this in conjunction with the increase of members in ENDS-specific subreddits may reflect the findings of Dai and Leventhal that reported increases in current and daily ENDS use among young adult populations between 2014 and 2018 (63). However, we observed drastic decreases in both the number of members and number of posts within ENDS-specific subreddits in 2018. This observation contradicts survey findings that the prevalence of ENDS use among US adults in 2018 had reached a watermark high of 7.6% over the previous 5 years (63). Most posts made in these subreddits focused on nicotine flavors, accessories, procurement of ENDS devices, as

⁹<https://www.statista.com/statistics/261766/share-of-us-internet-users-who-use-reddit-by-age-group/>




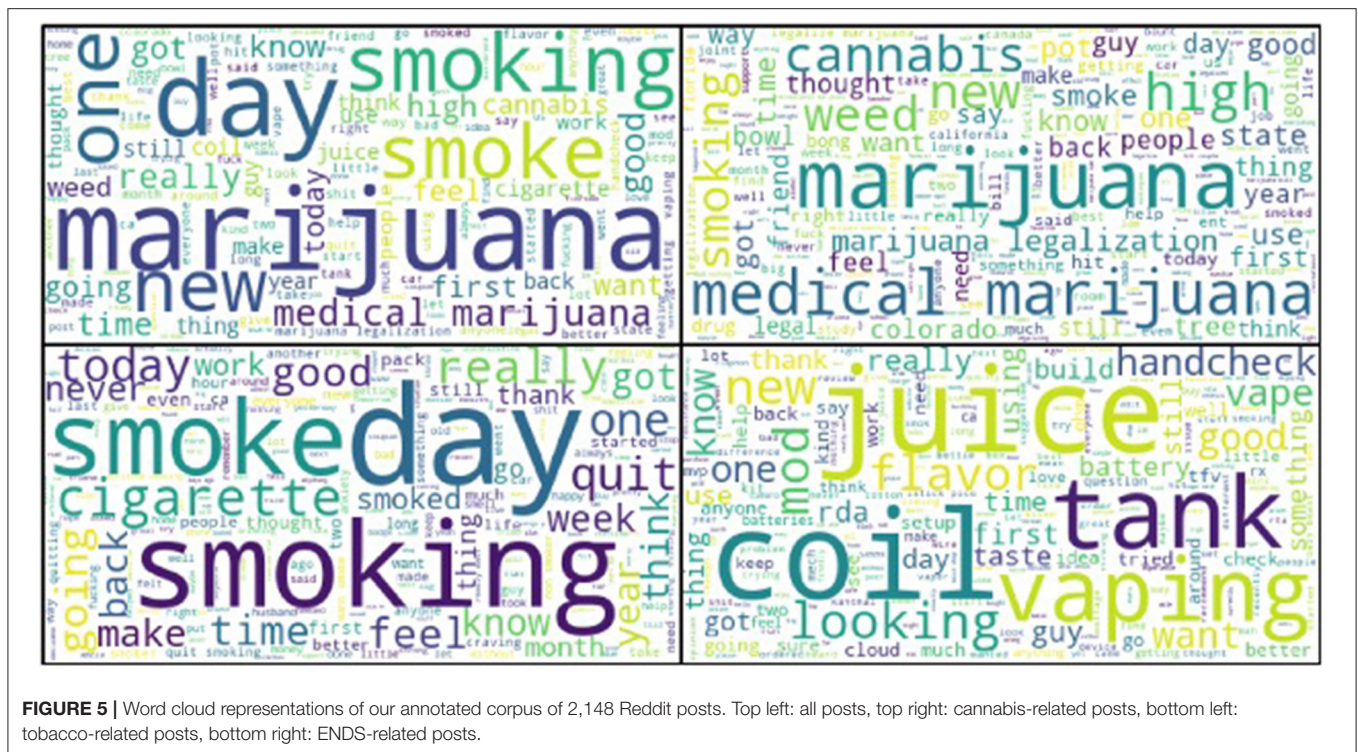
Product	LDA Topic Label	Synthetic Example
 <p>Cannabis</p>	<p>Personal Experiences (“feel”, “high”, “friends”, “job”, “favorite”, “test”, “happy”)</p> <p>Forms of Use (“kush”, “flower”, “wax”, “butter”, “oil”, “bud”, “dab”)</p> <p>Legalization/Regulation (“medical”, “legal”, “bill”, “dispensary”, “card”, “recreational”)</p> <p>Health Effect/Uses (“anxiety”, “pain”, “cbd”, “feel”, “sleep”, “heart”, “effects”)</p> <p>Cultivation (“growing”, “plant”, “seeds”, “quality”, “marijuana”)</p> <p>Methods of Use (“piece”, “bong”, “rig”, “pipe”, “bubbler”, “glass”)</p>	<p>→ “I got high as a kite at work last night.”</p> <p>→ “Oil gets me way higher than bud or wax.”</p> <p>→ “A new bill is being drafted to legalize in my city.”</p> <p>→ “Smoking has helped my pain and anxiety so much.”</p> <p>→ “Decided to try growing my own pot plants.”</p> <p>→ “I will take a bong over a pipe all day everyday!”</p>
 <p>ENDS</p>	<p>Flavors (“tank”, “juice”, “burnt”, “flavor”, “taste”)</p> <p>Accessories/Parts (“battery”, “tank”, “wire”, “charger”, “coil”, “wattage”)</p> <p>Smoking Cessation (“mg”, “nicotine”, “smoking”, “quit”, “help”, “vaping”, “want”)</p> <p>Purchasing/Acquisition (“order”, “shipping”, “get”, “new”, “online”, “customer”, “order”)</p>	<p>→ “Has anyone else had vape juice that tastes burnt?”</p> <p>→ “My new mech mod runs at 70 watts. Do you even vape bro?”</p> <p>→ “Officially one month smoke free thanks to my vape!”</p> <p>→ “My new mod shipped and is supposed to get here today.”</p>
 <p>Tobacco</p>	<p>Brands and Types (“Camel”, “menthol”, “Reds”, “American”, “Marlboro”, “Turkish”)</p> <p>Smoking Cessation (“cravings”, “quit”, “help”, “without”, “anxiety”, “patches”, “vaping”)</p>	<p>→ “What is everyones go to? Mine is Marlboro Reds”</p> <p>→ “Day 3 without a smoke and I still don’t want one puff.”</p>

FIGURE 4 | Resulting topics from our LDA analysis, accompanied by common terms observed in the topic and synthetic textual examples.



well as smoking cessation using these devices. Reddit members in the *r/Vaping* and *r/electronic_cigarette* subreddits often posted about their use patterns and experiences with ENDS devices. Consequently, member-specific participation in ENDS-related subreddits, as well as their participation in separate health-related subreddits over time (e.g., *r/Health*), may provide insight and generate hypotheses into the potential health effects attributable to prolonged ENDS use, and serves as a potential avenue for future computational epidemiology research.

As the scope of our manual annotation (i.e., characterizing product mentions of cannabis, tobacco, and ENDS) and our topic modeling differed, we developed word clouds in an attempt to broadly compare the similarity of discourse between posts contained in our annotated corpus and the larger macro-corpus. We found that the frequently observed terms in our annotated corpus closely resembled those topics discovered in our topic modeling, and therefore we hypothesize that many of the characteristics seen in our annotated corpus (e.g., members frequently discussing present cannabis use patterns, discussion of present smoking cessation attempts) may be frequently discussed throughout the entire macro-corpus, providing a potential avenue for future work.

In our manual annotation, we only observed three explicitly mentioned instances of product dual use. This is likely an artifact of our sampling strategy as we only annotated posts from 124 specific members where dual use was evidently uncommon. We anticipated more instances of product dual use than we observed in our manual annotation, as a substantial body of epidemiologic studies have demonstrated frequent use of multiple product types (1–4), and through the observation that many members posted in multiple subreddits as seen in **Figure 2**. Though we

were not able to ascertain frequent dual use within this study, Reddit members may disclose these habits within individual posts, outside of the relatively small sample size of annotated posts reported in this study. Few posts explicitly discussed transitions between cannabis, ENDS, and tobacco. And of the posts that did mention transitions, all of said posts discussed transitions between ENDS and combustible tobacco use, with no posts mentioning transitions to or from cannabis. These posts frequently discussed current smoking cessation attempts facilitated by ENDS, in addition to relapse events from ENDS devices to combustible tobacco. These findings showed that Reddit members in our annotated corpus seldom talked about transitions between these products, and when they did, they were often centered around usage in smoking cessation attempts. However, these patterns of transition may be more common in the larger macro-corpus upon further analysis.

Our study has some limitations to be considered. First, our analysis was limited to posts from eight popular subreddits focused on cannabis, ENDS, and tobacco. However, the content posted in these subreddits may contain posts discussing non-related topics (i.e., topics not related to cannabis, ENDS, tobacco). Although our work focused on larger subreddits, additional subreddits also discuss cannabis, ENDS, and tobacco products, such as *r/leaves* (i.e., discussion of cessation from THC-containing products). Future work may look to build on our analysis and include these additional subreddits closely related to cannabis, ENDS, and tobacco products (e.g., *r/cannabis*, *r/vaparents*, *r/DIY_eJuice*, *r/cigars*, *r/hookah*). Second, while we manually annotated 2,630 Reddit posts, these posts only encompassed 124 Reddit members as a result of our sampling strategy. Although we observed use patterns comparable to

prior epidemiologic studies, we cannot ascertain the geographic locations of Reddit members, nor can we assume that posts, opinions, and use patterns observed in our analysis are representative of all cannabis, ENDS, and tobacco members. Therefore, we cannot generalize these findings to the general population. Furthermore, this small sample of members may not encompass the true representation of our attributes of interest as seen on Reddit. Third, social media sites often fall victim to a phenomenon known as the 90-9-1 principle (64), where the large majority of members on a social media site just observe posts and do not contribute, a small proportion of members contribute sparingly, and a small number of members who contribute the majority of posts. Consequently, any findings resulting from social media data are difficult to generalize to the general population. Fourth, as seen in **Figure 1**, there was an observed decrease in posts collected during 2018. We expected to see continual increases in posts made in 2018, as a result, these observed decreases may be a result of decreases in the number of posts retrievable by the API, or discrepancies with parsing posts collected by the API. Finally, the posts in our analysis were collected between 2013 and 2018. With the consistently changing landscape regarding ENDS devices, cannabis legislation, and combustible tobacco use, our results may differ upon the analysis of more recent posts from these subreddits.

CONCLUSION

In conclusion, our study compared aspects of cannabis, ENDS, and tobacco use across multiple subreddits on the social media site Reddit. We found that Reddit posts provide firsthand accounts of cannabis, ENDS, and tobacco use and can complement findings derived from traditional survey-based approaches. In subreddits dedicated to cannabis, members frequently discussed personal experiences, methods and forms of use, legislation and policy, the associated health effects of its use, as well as the cultivation of cannabis plants. In tobacco specific subreddits, members often talked about ENDS devices, documented their smoking cessation attempts, and discussed brands and types of cigarettes. Further, ENDS-specific subreddits were often focused on parts and accessories, flavors and nicotine solutions, the procurement of ENDS devices, and discussion of smoking cessation using ENDS devices. Upon closer examination of these posts, the *r/trees*, *r/stopsmoking*, *r/Vaping*, and *r/electronic_cigarette* subreddits were more commonly used to discuss personal experiences with cannabis, ENDS, and tobacco. Future computational research should look to expand upon the manual annotation scheme presented and develop models to

classify personal experiences from general product mentions for a more focused analysis on cannabis, ENDS, and tobacco use as presented on Reddit.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because this dataset contains social media usernames which if shared, may result in the identification of those individuals. Consequently, this data is not to be shared. Requests to access the datasets should be directed to ryzen.benson@utah.edu.

AUTHOR CONTRIBUTIONS

RB and MC conceived the study idea with support from S-HZ. MC, MH, AC, and RB developed the annotation scheme presented and carried out the manual annotation of the data. RB developed the computational models and analyzed the data. RB also led the manuscript writing with support from MC, AC, and S-HZ. All authors contributed to the article and approved the submitted version.

FUNDING

The research reported in this publication was partially supported by the National Institute on Drug Abuse of the National Institutes of Health under award number R21DA043775. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the National Institute on Drug Abuse. This study was also partially supported by grant number T15LM007124 from the National Library of Medicine. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Library of Medicine or the National Institutes of Health.

ACKNOWLEDGMENTS

We would like to sincerely thank Sheetal Hardikar PhD, Greg Stoddard MPH, and Tengda Lin MPH at the University of Utah for their guidance and insight into the statistical analyses of our findings.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpubh.2021.738513/full#supplementary-material>

REFERENCES

- Wills TA, Knight R, Williams RJ, Pagano I, Sargent JD. Risk factors for exclusive E-cigarette use and dual E-cigarette use and tobacco use in adolescents. *Pediatrics*. (2015) 135:e43–51. doi: 10.1542/peds.2014-0760
- Berg CJ, Stratton E, Schauer GL, Lewis M, Wang Y, Windle M, et al. Perceived harm, addictiveness, and social acceptability of tobacco products and marijuana among young adults: marijuana, hookah, and electronic cigarettes win. *Subst Use Misuse*. (2015) 50:79–89. doi: 10.3109/10826084.2014.958857
- Azagba S. E-cigarette use, dual use of e-cigarettes and tobacco cigarettes, and frequency of cannabis use among high school students. *Addict Behav*. (2018) 79:166–70. doi: 10.1016/j.addbeh.2017.12.028
- Cohn AM, Abudayyeh H, Perreras L, Peters EN. Patterns and correlates of the co-use of marijuana with any tobacco and individual tobacco products in

- young adults from Wave 2 of the PATH study. *Addict Behav.* (2019) 92:122–7. doi: 10.1016/j.addbeh.2018.12.025
5. McDonald EA, Popova L, Ling PM. Traversing the triangulum: the intersection of tobacco, legalised marijuana and electronic vaporisers in Denver, Colorado. *Tob Control.* (2016) 25:i96–102. doi: 10.1136/tobaccocontrol-2016-053091
 6. United States Public Health Service Office of the Surgeon General, National Center for Chronic Disease Prevention and Health Promotion (US) Office on Smoking and Health. *Smoking Cessation: A Report of the Surgeon General.* Washington, DC: US Department of Health and Human Services (2020). Available online at: <http://www.ncbi.nlm.nih.gov/books/NBK555591/> (accessed January 13, 2021).
 7. Cornelius ME, Wang TW, Jamal A, Loretan CG, Neff LJ. Tobacco product use among adults — United States, 2019. *MMWR Morb Mortal Wkly Rep.* (2020) 69:1736–42. doi: 10.15585/mmwr.mm6946a4
 8. National Center for Chronic Disease Prevention and Health Promotion (US) Office on Smoking and Health. *The Health Consequences of Smoking – 50 Years of Progress: A Report of the Surgeon General: (510072014-001).* Atlanta, GA: Centers for Disease Control and Prevention (US) (2014). Available online at: <http://www.ncbi.nlm.nih.gov/books/NBK179276/> (accessed December 1, 2021).
 9. Babb S. Quitting smoking among adults — United States, 2000–2015. *MMWR Morb Mortal Wkly Rep.* (2017) 65:1457–64. doi: 10.15585/mmwr.mm6552a1
 10. Franck C, Budlovsky T, Windle SB, Fillion KB, Eisenberg MJ. Electronic cigarettes in North America: history, use, and implications for smoking cessation. *Circulation.* (2014) 129:1945–52. doi: 10.1161/CIRCULATIONAHA.113.006416
 11. Bullen C, Howe C, Laugesen M, McRobbie H, Parag V, Williman J, et al. Electronic cigarettes for smoking cessation: a randomised controlled trial. *Lancet.* (2013) 382:1629–37. doi: 10.1016/S0140-6736(13)61842-5
 12. Masiero M, Lucchiari C, Mazzocco K, Veronesi G, Maisonneuve P, Jemos C, et al. E-cigarettes may support smokers with high smoking-related risk awareness to stop smoking in the short run: preliminary results by randomized controlled trial. *Nicot Tobacco Res.* (2019) 21:119–26. doi: 10.1093/ntr/nty047
 13. Hajek P, Phillips-Waller A, Przulj D, Pesola F, Myers Smith K, Bisal N, et al. A randomized trial of E-cigarettes versus nicotine-replacement therapy. *N Engl J Med.* (2019) 380:629–37. doi: 10.1056/NEJMoa1808779
 14. Zhu S-H, Zhuang Y-L, Wong S, Cummins SE, Tedeschi GJ. E-cigarette use and associated changes in population smoking cessation: evidence from US current population surveys. *BMJ.* (2017) 358:j3262. doi: 10.1136/bmj.j3262
 15. Creamer MR, Everett Jones S, Gentzke AS, Jamal A, King BA. Tobacco product use among high school students — youth risk behavior survey, United States, 2019. *MMWR Suppl.* (2020) 69:56–63. doi: 10.15585/mmwr.su6901a7
 16. Wang TW. E-cigarette use among middle and high school students — United States, 2020. *MMWR Morb Mortal Wkly Rep.* (2020) 69:1310–2. doi: 10.15585/mmwr.mm6937e1
 17. Henningfield J, Pankow J, Garrett B. Ammonia and other chemical base tobacco additives and cigarette nicotine delivery: issues and research needs. *Nicotine Tob Res.* (2004) 6:199–205. doi: 10.1080/1462220042000202472
 18. O’Connell G, Pritchard JD, Prue C, Thompson J, Verron T, Graff D, et al. A randomised, open-label, cross-over clinical study to evaluate the pharmacokinetic profiles of cigarettes and e-cigarettes with nicotine salt formulations in US adult smokers. *Intern Emerg Med.* (2019) 14:853–61. doi: 10.1007/s11739-019-02025-3
 19. Benson R, Hu M, Chen AT, Nag S, Zhu S-H, Conway M. Investigating the attitudes of adolescents and young adults towards JUUL: computational study using twitter data. *JMIR Public Health Surveill.* (2020) 6:e19975. doi: 10.2196/19975
 20. Vallone DM, Bennett M, Xiao H, Pitzer L, Hair EC. Prevalence and correlates of JUUL use among a national sample of youth and young adults. *Tob Control.* (2019) 28:603–9. doi: 10.1136/tobaccocontrol-2018-054693
 21. Evans-Polce R, Veliz P, Boyd CJ, McCabe VV, McCabe SE. Trends in E-cigarette, cigarette, cigar, and smokeless tobacco use among US adolescent cohorts, 2014–2018. *Am J Public Health.* (2020) 110:163–5. doi: 10.2105/AJPH.2019.305421
 22. Dobbs PD, Hodges EJ, Dunlap CM, Cheney MK. Addiction vs. dependence: a mixed methods analysis of young adult JUUL users. *Addict Behav.* (2020) 107:106402. doi: 10.1016/j.addbeh.2020.106402
 23. Case KR, Hinds JT, Creamer MR, Loukas A, Perry CL. Who is JUULing and why? An examination of young adult electronic nicotine delivery systems users. *J Adolesc Health.* (2020) 66:48–55. doi: 10.1016/j.jadohealth.2019.05.030
 24. Soneji S, Barrington-Trimis JL, Wills TA, Leventhal AM, Unger JB, Gibson LA, et al. Association between initial use of e-cigarettes and subsequent cigarette smoking among adolescents and young adults: a systematic review and meta-analysis. *JAMA Pediatr.* (2017) 171:788–97. doi: 10.1001/jamapediatrics.2017.1488
 25. Whiting PE, Wolff RF, Deshpande S, Di Nisio M, Duffy S, Hernandez AV, et al. Cannabinoids for medical use: a systematic review and meta-analysis. *JAMA.* (2015) 313:2456–73. doi: 10.1001/jama.2015.6358
 26. Memedovich KA, Dowsett LE, Spackman E, Noseworthy T, Clement F. The adverse health effects and harms related to marijuana use: an overview review. *CMAJ Open.* (2018) 6:E339–46. doi: 10.9778/cmajo.20180023
 27. Azofeifa A, Mattson ME, Schauer G, McAfee T, Grant A, Lyerla R. National estimates of marijuana use and related indicators — national survey on drug use and health, United States, 2002–2014. *Morb Mort Wkly Rep.* (2016) 65:1–25. doi: 10.15585/mmwr.ss6511a1
 28. Goodman S, Wadsworth E, Leos-Toro C, Hammond D. Prevalence and forms of cannabis use in legal vs. illegal recreational cannabis markets. *Int J Drug Policy.* (2020) 76:102658. doi: 10.1016/j.drugpo.2019.102658
 29. Hasin DS, Sarvet AL, Cerdá M, Keyes KM, Stohl M, Galea S, et al. US adult illicit cannabis use, cannabis use disorder, and medical marijuana laws: 1991–1992 to 2012–2013. *JAMA Psychiatry.* (2017) 74:579–88. doi: 10.1001/jamapsychiatry.2017.0724
 30. ElSohly MA, Mehmedic Z, Foster S, Gon C, Chandra S, Church JC. Changes in cannabis potency over the last two decades (1995–2014) - analysis of current data in the United States. *Biol Psychiatry.* (2016) 79:613–9. doi: 10.1016/j.biopsych.2016.01.004
 31. Steigerwald S, Wong PO, Khorasani A, Keyhani S. The form and content of cannabis products in the United States. *J Gen Intern Med.* (2018) 33:1426–8. doi: 10.1007/s11606-018-4480-0
 32. Park A, Conway M. Tracking health related discussions on reddit for public health applications. *AMIA Annu Symp Proc.* (2018) 2017:1362–71.
 33. Bunting AM, Frank D, Arshonsky J, Bragg MA, Friedman SR, Krawczyk N. Socially-supportive norms and mutual aid of people who use opioids: an analysis of Reddit during the initial COVID-19 pandemic. *Drug Alcohol Depend.* (2021) 222:108672. doi: 10.1016/j.drugalcdep.2021.108672
 34. Krawczyk N, Bunting AM, Frank D, Arshonsky J, Gu Y, Friedman SR, et al. “How will I get my next week’s script?” Reactions of Reddit opioid forum users to changes in treatment access in the early months of the coronavirus pandemic. *Int J Drug Policy.* (2021) 92:103140. doi: 10.1016/j.drugpo.2021.103140
 35. Meacham MC, Paul MJ, Ramo DE. Understanding emerging forms of cannabis use through an online cannabis community: an analysis of relative post volume and subjective highness ratings. *Drug Alcohol Depend.* (2018) 188:364–9. doi: 10.1016/j.drugalcdep.2018.03.041
 36. Luo J, Chen L, Lu X, Yuan J, Xie Z, Li D. Analysis of potential associations of JUUL flavours with health symptoms based on user-generated data from Reddit. *Tob Control.* (2021) 30:534–41. doi: 10.1136/tobaccocontrol-2019-055439
 37. Tamersoy A, De Choudhury M, Chau DH. Characterizing smoking and drinking abstinence from social media. *HT ACM Conf Hypertext Soc Media.* (2015) 2015:139–48. doi: 10.1145/2700171.2791247
 38. Hu M, Benson R, Chen AT, Zhu S-H, Conway M. Determining the prevalence of cannabis, tobacco, and vaping device mentions in online communities using natural language processing. *Drug Alcohol Depend.* (2021) 228:109016. doi: 10.1016/j.drugalcdep.2021.109016
 39. Baumgartner JM. *pushshift/api.* (2021). Available online at: <https://github.com/pushshift/api> (accessed January 14, 2021).
 40. Baumgartner J, Zannettou S, Keegan B, Squire M, Blackburn J. *The Pushshift Reddit Dataset. arXiv:200108435 [cs].* (2020). Available online at: <http://arxiv.org/abs/2001.08435> (accessed February 10, 2021).

41. Park A, Hartzler AL, Huh J, Hsieh G, McDonald DW, Pratt W. "How Did We Get Here?": topic drift in online health discussions. *J Med Internet Res.* (2016) 18:e284. doi: 10.2196/jmir.6297
42. Myslín M, Zhu S-H, Chapman W, Conway M. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. *J Med Internet Res.* (2013) 15:e174. doi: 10.2196/jmir.2534
43. Mowery D, South B, Patterson O, Zhu S-H, Conway M. Investigating the documentation of electronic cigarette use in the veteran affairs electronic health record: a pilot study. In: *BioNLP 2017*. Vancouver, BC: Association for Computational Linguistics (2017). p. 282–6. doi: 10.18653/v1/W17-2335
44. Zhu S-H, Sun JY, Bonnevie E, Cummins SE, Gamst A, Yin L, et al. Four hundred and sixty brands of e-cigarettes and counting: implications for product regulation. *Tob Control.* (2014) 23:iii3–9. doi: 10.1136/tobaccocontrol-2014-051670
45. Rehurek R. *gensim: Python Framework for Fast Vector Space Modelling*. Available online at: <http://radimrehurek.com/gensim> (accessed January 15, 2021).
46. South B, Shen S, Leng J, Forbush T, DuVall S, Chapman W. A prototype tool set to support machine-assisted annotation. In: *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*. Montréal, QC: Association for Computational Linguistics (2012). p. 130–9. Available online at: <https://aclanthology.org/W12-2416> (accessed December 1, 2021).
47. Hripacsak G, Rothschild AS. Agreement, the F-measure, and reliability in information retrieval. *J Am Med Inform Assoc.* (2005) 12:296–8. doi: 10.1197/jamia.M1733
48. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Measur.* (1960) 20:37–46. doi: 10.1177/001316446002000104
49. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med.* (2012) 22:276–82. doi: 10.11613/BM.2012.031
50. Bird S, Klein E, Loper E. *Natural Language Processing with Python*. 1st ed. Beijing; Cambridge, MA: O'Reilly (2009).
51. Bengfort B, Bilbro R, Ojeda T. *Applied Text Analysis With Python: Enabling Language-Aware Data Products with Machine Learning*. 1st ed. Sebastopol, CA: O'Reilly Media, Inc. (2018).
52. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *J Mach Learn Res.* (2003) 3:993–1022. doi: 10.5555/944919.944937
53. Müller AC, Guido S. *Introduction to Machine Learning With Python: A Guide for Data Scientists*. 1st ed. Sebastopol, CA: O'Reilly Media, Inc. (2016).
54. Ghosh D, Guha R. What are we 'tweeting' about obesity? Mapping tweets with topic modeling and geographic information system. *Cartogr Geogr Inf Sci.* (2013) 40:90–102. doi: 10.1080/15230406.2013.776210
55. Surian D, Nguyen DQ, Kennedy G, Johnson M, Coiera E, Dunn AG. Characterizing twitter discussions about HPV vaccines using topic modeling and community detection. *J Med Internet Res.* (2016) 18:e232. doi: 10.2196/jmir.6045
56. Liu Q, Zheng Z, Zheng J, Chen Q, Liu G, Chen S, et al. Health communication through news media during the early stage of the COVID-19 outbreak in china: digital topic modeling approach. *J Med Internet Res.* (2020) 22:e19118. doi: 10.2196/19118
57. Mabe B. *pyLDAvis: Interactive Topic Model Visualization*. Port of the R package. Available online at: <https://github.com/bmabey/pyLDAvis> (accessed January 17, 2021).
58. Mueller A. *amueller/word_cloud*. (2021). Available online at: https://github.com/amueller/word_cloud (accessed January 15, 2021).
59. Cavazos-Rehg PA, Krauss M, Fisher SL, Salyer P, Gruzca RA, Bierut LJ. Twitter chatter about marijuana. *J Adolesc Health.* (2015) 56:139–45. doi: 10.1016/j.jadohealth.2014.10.270
60. Daniulaityte R, Nahhas RW, Wijeratne S, Carlson RG, Lamy FR, Martins SS, et al. "Time for dabs": analyzing Twitter data on marijuana concentrates across the U.S. *Drug Alcohol Depend.* (2015) 155:307–11. doi: 10.1016/j.drugalcdep.2015.07.1199
61. General USPHSO of the S, Health NC for CDP and HP (US) O on S and. *Patterns of Smoking Cessation Among U.S. Adults, Young Adults, and Youth*. US Department of Health and Human Services (2020). Available online at: <https://www.ncbi.nlm.nih.gov/books/NBK555598/> (accessed December 15, 2020).
62. Chen AT, Zhu S-H, Conway M. What online communities can tell us about electronic cigarettes and hookah use: a study using text mining and visualization techniques. *J Med Internet Res.* (2015) 17:e220. doi: 10.2196/jmir.4517
63. Dai H, Leventhal AM. Prevalence of e-cigarette use among adults in the United States, 2014–2018. *JAMA.* (2019) 322:1824. doi: 10.1001/jama.2019.15331
64. van Mierlo T. The 1% rule in four digital health social networks: an observational study. *J Med Internet Res.* (2014) 16:e33. doi: 10.2196/jmir.2966

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Benson, Hu, Chen, Zhu and Conway. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.