Title

Determining the prevalence of cannabis, tobacco, and vaping device mentions in online communities using natural language processing

Authors

Hu, Mengke
Benson, Ryzen
Chen, Annie T
et al.

Publication Date

2021-11-01

DOI

10.1016/j.drugalcdep.2021.109016

Peer reviewed

Determining the prevalence of cannabis, tobacco, and vaping device mentions in online communities using natural language processing

Mengke Hu*

Department of Biomedical Informatics, University of Utah, Salt Lake City, UT

Ryzen Benson

Department of Biomedical Informatics, University of Utah, Salt Lake City, UT

Annie T. Chen

Department of Biomedical Informatics & Medical Education, University of Washington, Seattle, WA

Shu-Hong Zhu

Herbert Wertheim School of Public Health, University of California San Diego, La Jolla, CA

Mike Conway

Department of Biomedical Informatics, University of Utah, Salt Lake City, UT

Abstract

Introduction. The relationship between cannabis, tobacco, and vaping devices is both rapidly changing and poorly understood, with consumers rapidly shifting between use of all three product types. Given this dynamic and evolving landscape, there is an urgent need to monitor and better understand co-use, dual-use, and transition patterns between these products. This study describes work that utilizes social media — in this case, Reddit — in

[*] Corresponding author

conjunction with automated **N**atural **L**anguage **P**rocessing (**NLP**) methods to better understand cannabis, tobacco, and vaping device product usage patterns.

Methods. We collected Reddit data from the period 2013-2018 sourced from eight popular, high-volume Reddit communities (*subreddits*) related to the three product categories. We then manually annotated (*coded*) a set of 2,640 Reddit posts and trained a machine learning-based NLP algorithm to automatically identify and disambiguate between cannabis or tobacco mentions (both smoking and vaping) in Reddit posts. This classifier was then applied to all data derived from the eight subreddits, 767,788 posts in total.

Results. The NLP algorithm achieved an overall moderate performance (overall F-score of 0.77). When applied to our large corpus of Reddit posts, we discovered that over 10% of posts in the smoking cessation subreddit `r/stopsmoking` were classified as referring to vaping nicotine, and that only 2% of posts from the subreddits `r/electronic_cigarette` and `r/vaping` were classified as referring to smoking (tobacco) cessation.

Conclusions. This study presents the results of applying an NLP algorithm designed to identify and distinguish between cannabis and tobacco mentions (both smoking and vaping) in Reddit posts, hence contributing to our currently limited understanding of co-use, dual-use, and transition patterns between these products.

Keywords: cannabis, tobacco, social media, natural language processing

1. Introduction

Tobacco use — specifically cigarette smoking — is associated with an increased risk of heart disease, respiratory disease, various kinds of cancer, and general poor health (Warner, 2014). Overall, current smokers in the United States are estimated to lose over a decade of life when compared to never smokers, with over 480,000 deaths per year directly attributable to smoking (Surgeon General, 2014).

Electronic **N**icotine **D**elivery Systems (**ENDS**) were first introduced to the US market in 2007. Once established in the US, the product experienced exponential growth, with the number of ENDS users doubling every year between 2008 and 2012 (Grana et al., 2014). The risks — and potential benefits — of ENDS use are currently poorly understood (Surgeon General, 2020).

Cannabis is used by 7.4% of the US population between the ages of 12 and 17 (around 1.8 million individuals), rising to 20% of young adults (6.8 million individuals aged 18-25 (SAMHSA, 2015)). While there are several conditions for which cannabis is believed to have therapeutic potential (e.g. *chronic pain*, *epilepsy* (Breijyeh et al., 2021)), there exist known harms associated with the long term use of cannabis (e.g. *chronic bronchitis*, *mental health problems*) particularly among adolescents (Volkow et al., 2014).

Survey data suggests that there is a high degree of co-use between cannabis, tobacco, and vaping devices (McDonald et al., 2016). Further, it has been found that dual-users — as opposed to individuals who use just one substance — are characterized by demographic and behavioral differences, with regular ENDS users much more likely to choose a vaping device as a primary means of cannabis administration, rather than combustible cannabis (Smith et al., 2020). Survey data has demonstrated a strong bidirectional relationship between cannabis use and combustible tobacco use in young adults (Doran et al., 2019), with increases in cannabis use predicting increases in tobacco use over time. Perhaps the most widespread co-use practice is the use of *blunts* (i.e. a cigar or cigarillo emptied of tobacco and then stuffed with cannabis), with 13.4% of 12 to 25 year olds in the US having used blunts within the last 12 months (Delnevo et al., 2011).

Social media — here defined to include internet discussion forums like Reddit — is a useful resource for substance use research given that it provides a readily-accessible source of naturalistic, publicly accessible first person narratives with which to understand rapidly changing health behaviors and attitudes (Foufi et al., 2019; Paul and Dredze, 2018; Park and Conway, 2017; Wongkoblap et al., 2017), and as such is especially suitable for the task of analyzing behaviors and attitudes related to tobacco, cannabis, and vaping (Meacham et al., 2018; Chen et al., 2015; Czaplicki et al., 2019; Ayers et al., 2017; Myslín et al., 2013).

Various applications of automated social media analytic methods to address substance use-related research questions have been reported in the literature. Facebook data has been used to investigate public attitudes towards synthetic opioids (Beletsky et al., 2020), to discover the emotional valence regarding alcohol use among college students (Van Swol et al., 2020), and to investigate and characterize substance use among new mothers (Oram et al., 2018). In addition to Facebook, online communities (including Reddit) have shown utility as a data source for investigating substance use behaviour. For example, work on identifying smoking status of individuals in the context of a smoking cessation community (Wang et al., 2019), and assessing attitudes towards the use of medication-assisted treatment for treating substance use disorders (Tofighi et al., 2021). Further, there has been considerable work using Twitter specifically focused on surveilling public attitudes towards tobacco, cannabis, and vaping device use (Kim et al., 2020; Myslín et al., 2013; Emery et al., 2014; Benson et al., 2020).

Reddit, a social media platform that has surged in popularity in recent years, has been utilized as a source of data for population health research using **N**atural **L**anguage **P**rocessing (**NLP**[1]), including work on investigating mental health issues like depression (Gkotsis et al., 2017; Park and Conway, 2017), identifying adverse drug reactions (Nguyen et al., 2017; Sarker et al., 2016), and characterizing use patterns for various substances (Cavazos-Rehg et al., 2019; Tamersoy et al., 2015; Chen et al., 2015; Conway et al., 2019). Work on using Reddit data to study the product types that are the focus of this paper are — in general — less well-developed. Exceptions include investigations into the frequency with which new forms of cannabis product consumption are discussed on Reddit (Meacham et al., 2018), an exploration of experiences and attitudes related to hookah, vaping, and combustible to-

[1] Natural Language Processing is, broadly speaking, the automated conversion of unstructured textual data to a structured format (Jurafsky and Martin, 2009). NLP approaches to processing social media data are advantageous due to their scalability, with automated algorithms capable of automatically analyzing millions of social media posts, unlike qualitative content analysis approaches that are typically limited to hundreds — or at best, thousands — of posts.

bacco (Chen et al., 2015), and an analysis of vaping-related posts across Reddit (Barker and Rohde, 2019).

As Goffman (1963) noted, in the context of socially marginalized, stigmatized behavior or characteristics, individuals may be drawn to anonymous communities as a means of gaining support, camaraderie, and a sense of solidarity while maintaining privacy in other areas of their lives. A key advantage of using Reddit as a data source is that the platform is organized into distinct topic-focused communities (*subreddits*) that can provide ready access to otherwise hard-to-reach stigmatized groups (e.g. risky sexual behavior among individuals who misuse opioids (Cavazos-Rehg et al., 2019), pro-eating disorder advocates (Sowles et al., 2018)). Reddit users access the platform using pseudo-anonymous usernames, and this — in addition to the availability of "throwaway" usernames — may encourage users to openly discuss sensitive topics that they may not be willing to disclose in the context of a survey or interview (Grucza et al., 2007; Amaya et al., 2019). Empirical support for this contention is supplied by Correa et al. (2015), who shows that users of anonymous social media platforms are more likely to articulate highly personal information (wants, needs, wishes, fears) than users of social media platforms that do not provide anonymity.

In addition to its advantages, Reddit — like all data sources — has limitations. First, geolocation information is not available, and hence we cannot ascertain with certainty in which country any given user resides (although approximately half the traffic to the service has an origin in the US (Tankovska, 2021)). Second, compared with other sources of big data (e.g. sensor data derived from wearable devices (Knotta et al., 2021)) Reddit data is primarily unstructured and requires NLP methods to extract meaning at scale.

There are many subreddits that focus on cannabis, tobacco, and vaping devices. Given our research goal — i.e. identifying usage patterns between the product groups — we selected eight high volume subreddits (i.e. subreddits with more than 50,000 members) that are explicitly focused on one of the three products, allowing us to address questions such as the following: *what proportion of posts in smoking cessation subreddits discuss vaping nicotine?*; *what proportion of posts in the vaping subreddits discuss cessation?*

With this broad context in mind, our more granular goals for this research were as follows. First, to annotate product mentions in a subset of 2,650 Reddit posts pertaining to the three product types. Second, use this annotated corpus to train several NLP classifiers to automatically identify product use mentions over our corpus of 767,788 initiating posts, thus allowing us to quantify the extent to which discussion of multiple products occur in product-specific subreddits.

2. Materials and Methods

Our methods consisted of the following steps. First, we collected a large corpus of Reddit data, and then developed and validated an annotation scheme designed to label product mentions (cannabis, tobacco, and vaping devices). Second, we developed an annotated corpus of 2,650 posts based on this annotation scheme. Third, using our annotated corpus, we trained an NLP algorithm to resolve ambiguous terms such as “smoking” and “vaping”. Fourth, we applied this trained classifier to our large, unlabelled corpus of Reddit data to discover prevalence of product discussion in each subreddit.

Finally, we estimated the extent to which discussion of multiple products is evident in product-specific subreddits.

2.1. Reddit Data Collection and Annotation

We collected data from eight subreddits related to cannabis, tobacco, vaping devices, and smoking cessation (`r/vaping`, `r/electronic_cigarettes`, `r/vaping101`, `r/weed`, `r/trees`, `r/marijuana`, `r/stopsmoking`, and `r/cigarettes`). We included a smoking cessation subreddit (`r/stopsmoking`) as prior work has suggested that smoking cessation is a common motivation for users to switch to vaping devices (Chen et al., 2015; Wadsworth et al., 2016), despite mixed evidence regarding the efficacy of this approach (Hajek et al., 2019; Bhatnagar et al., 2019).

Data was collected using the `pushshift.io` **A**pplication **P**rogramming **I**nterface (**API**) (Pushshift, 2021). We used `pushshift.io` as our data source as pilot work demonstrated that harvesting data using `pushshift` yielded a more complete dataset than other Reddit data collection methods (Gaffney and Matias, 2018). Our final dataset consisted of a total of 767,788 submissions (initiating posts) and their associated comments (6,877,693 posts) from January 1, 2013 to December 31, 2018 derived from the eight subreddits. We binned the eight subreddits into four groups for analysis: *vaping*, *tobacco*, *cannabis*, and *smoking cessation*.

As users typically present their own experiences in a thread's initiating post (MacLean et al., 2015), with subsequent posts within a thread frequently subject to off-topic discussion (i.e. "topic drift"), we restrict our product-usage annotation and analysis to initiating posts. Previous work (Park et al., 2016; MacLean et al., 2015) suggests that initiating posts contain more autobiographical content related to the user's own health behavior and substance use practices (i.e. substance used and number of days without use) than subsequent, responding posts.

Using textual data derived from the eight subreddits listed above, we developed an annotation scheme — i.e. a coding scheme — to represent mentions of all three product types (see **Table 1**). In this annotation scheme, each word or phrase can be labelled with multiple concepts. For example, the phrase 'I smoke blunts{COMBUSTIBLE_TOBACCO & COMBUSTIBLE_CANNABIS} because the weed{COMBUSTIBLE_CANNABIS} where I live tastes terrible' refers to both combustible cannabis and combustible tobacco use. Annotation was performed by authors MC, RB, and AC using the `eHost`[2] annotation tool. Inter-rater agreement was 0.83 (F-score[3]), indicating strong agreement. Using this annotation scheme, we (authors MC and RB) then went on to annotate 2,650 posts derived from a (stratified) random sample of 124 Reddit users. Further details regarding the annotation process, and the sampling method are provided in supplementary materials.

2.2. Classification: Extracting Product Mentions

Using the annotated corpus described above, we trained NLP classifiers to automatically identify and label product use mentions with a view to quantifying the volume of product-related discussion in different subreddits. We used three classification algorithms. First, a rule-based algorithm was developed to disambiguate between vaping cannabis and vaping nicotine,

[2] `https://github.com/jianlins/ehost`

[3] Note that F-score has been shown to function as an effective surrogate for Cohen's kappa (Hripcsak and Rothschild, 2005)

Annotation Class	**Description**
VAPING_ CANNABIS	Reference to vaping cannabis (including dabbing)
VAPING_NICOTINE	Reference to vaping nicotine
COMBUST_CANNABIS	Reference to combustible cannabis consumption (not edibles)
COMBUST_TOBACCO	Reference to combustible tobacco use
SMOKING_CESSATION	Reference to quitting combustible tobacco use
BRAND	Reference to tobacco, cannabis, or vaping device brand

Table 1: Annotation scheme

and smoking cannabis and smoking nicotine, with rules implemented using regular expressions[4]. Rules were developed using intuition, experience, and the advice of a tobacco control researcher (author SZ).

Second, we used a **C**onditional **R**andom **F**ield (**CRF**) algorithm (Lafferty et al., 2001). CRFs are a popular method for sequence-based labeling (Chatzis and Demiris, 2013), and have been widely used for language processing applications. We used the `CRFsuite` (Okazaki, 2016) implementation of CRFs for this work.

Third, we utilized a bidirectional **L**ong **S**hort-**T**erm **M**emory **R**ecurrent **N**eural **N**etwork (**LSTM-RNN**) classifier (Hochreiter and Schmidhuber, 1997). LSTM-RNNs are well-suited to the task of identifying and disambiguating terms and phrases due to their sensitivity to long-range linguistic dependencies (i.e. context). Like CRFs, LSTM-RNNs have been widely used

[4] Regular expressions formally define a search pattern that can be used for pattern matching in text. For example, the regular expression `\bvap(e|er|or|our|ing)\b` matches the terms *vape*, *vaper*, *vapor*, *vapour*, and *vaping*. Regular expressions can be chained together to create complex rules.

for sequence-based NLP tasks in recent years, (Huang et al., 2015; Yang et al., 2019) We used `TensorFlow` (TensorFlow, 2019) to construct this network. A more comprehensive description of the technical approach adopted is available in supplementary materials.

Our goal in evaluating the three different approaches to identifying product mentions was to determine if (and to what extent) LSTM-RNNs perform better than simpler, but more readily intelligible, rule-based methods.

2.3. Experiments & Evaluation Methods

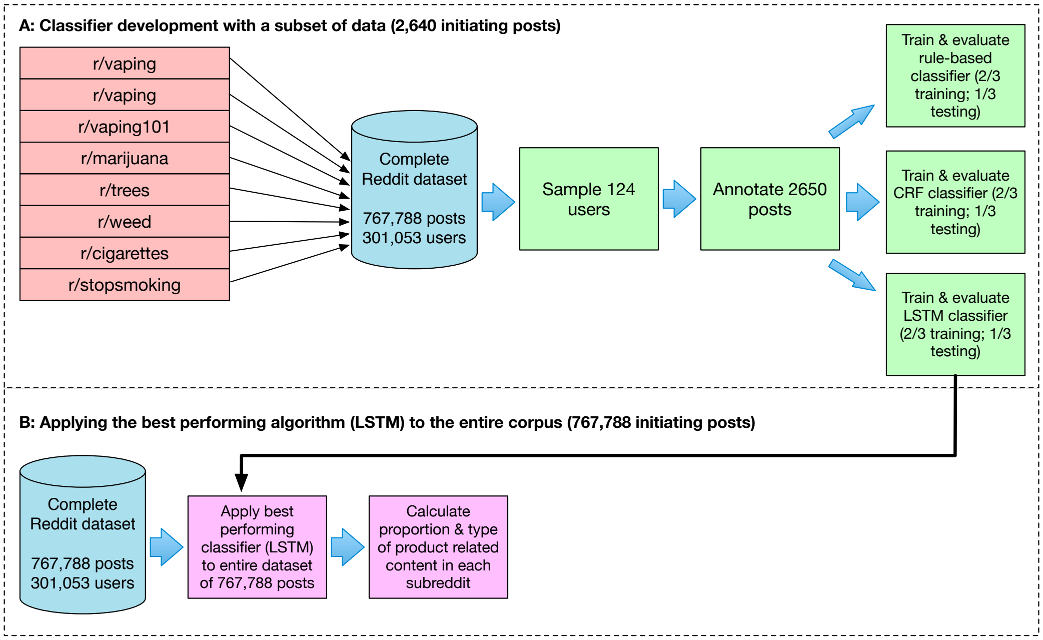
A: Classifier development with a subset of data (2,640 initiating posts)
r/vaping
r/vaping
r/vaping101
r/marijuana
r/trees
r/weed
r/cigarettes
r/stopsmoking
Complete Reddit dataset
767,788 posts
301,053 users
Sample 124 users
Annotate 2650 posts
Train & evaluate rule-based classifier (2/3 training; 1/3 testing)
Train & evaluate CRF classifier (2/3 training; 1/3 testing)
Train & evaluate LSTM classifier (2/3 training; 1/3 testing)
B: Applying the best performing algorithm (LSTM) to the entire corpus (767,788 initiating posts)
Complete Reddit dataset
767,788 posts
301,053 users
Apply best performing classifier (LSTM) to entire dataset of 767,788 posts
Calculate proportion & type of product related content in each subreddit

Figure 1: Experimental procedure

Our experimental procedure is described in **Figure 1**. To automatically label product mentions in text, we trained text classifiers using product mentions derived from $^2/_3$ of our annotated corpus (n = 1,764 posts), holding out product mentions derived from the remaining 1/3 (n = 886 posts) of our annotated corpus for testing. Performance was measured by calculating the overall precision (positive predictive value), recall (sensitivity), and F-score (harmonic mean of precision and recall) of the algorithm. This training and evaluation process is summarized in **Figure 1 (part A)**. After training and evaluation, we then went on to apply our best performing algorithm to our complete dataset of 767,788 initiating posts (i.e. the totality of initiating posts in all eight subreddits posted between 2013 and 2018). This process of applying the validated model to our entire dataset is summarized in **Figure 1 (part B)**. The experimental framework adopted in this study is standard practice in NLP research (Jurafsky and Martin, 2009).

2.4. Ethics Statement

This study was exempted from review by the University of Utah Institutional Review Board (IRB_00076188). In line with emerging ethical guidelines for social media research (Conway, 2014; Benton et al., 2017) and to protect user privacy, we have refrained from including usernames in this paper. Further, all quotations used are synthesized from multiple examples.

3. Results

3.1. Corpus Characteristics

Summary statistics for our larger corpus of Reddit data (i.e. all 767,788 initiating posts from the eight product-related subreddits), of which the annotated corpus is a small subset (i.e. 2,650 posts) are shown in **Table 2**, where it can be seen that the cannabis-related subreddit `r/trees` had the greatest number of users, the vaping related subreddit `r/electronic_cigarettes`

Subreddit	type	#users	#posts	Ave(#post/user)	post length(# words)
r/vaping	vaping	66,046	927,078	14.03	27.82
r/electronic_cigarettes	vaping	56,243	906,368	16.12	30.71
r/vaping101	vaping	1,578	5,366	3.40	48.18
r/weed	cannabis	4,760	11,447	2.40	21.93
r/trees	marijuana	468,580	4,191,113	8.94	23.71
r/marijuana	marijuana	40,261	239,409	5.95	32.82
r/stopsmoking	tobacco	57,297	550,559	9.61	50.45
r/cigarettes	tobacco	24,820	243,707	9.82	24.68

Table 2: Subreddit data from 2013 to 2018

had the largest number of posts per user, and members of the r/stopsmoking subreddit had the longest posts on average.

3.2. Classifier Performance

Of the three classifiers trained and tested on our annotated subset of data (i.e. rule-based, CRF, and LSTM-RNN), the LSTM-RNN-based method achieved the best overall F-score (0.77) over the six concept types listed (e.g. Vaping_Cannabis, Vaping_Nicotine, Combust_Cannabis, Combust_Tobacco, Smoking_Cessation and Brand) when applied to our annotated corpus, with the CRF algorithm performing substantially better (0.72 F-score) than the rule-based classifier (0.49 F-score, see **Table 3**). The classifiers were further evaluated at the concept level (i.e. the six concepts listed in **Table 1**). There were considerable differences between the classifiers' performance on the six categories, with results ranging from 0.24 (F-score) for Brand to 0.86 (F-score) for Combustible_Cannabis. However, in terms of F-score, the LSTM-RNN classifier consistently outperformed the

CRF and rule-based classifier for all categories.

3.3. Distribution of Product Mentions

	Rule-based			**CRF**			**LSTM-RNN**		
	PPV[1]	**Sens.**[2]	**F1**[3]	**PPV**	**Sens.**	**F1**	**PPV**	**Sens.**	**F1**
Vaping_Cannabis	0.08	0.08	0.08	0.86	0.08	0.15	0.71	0.21	0.32
Vaping_Nic/Tobacco	0.81	0.46	0.59	0.82	0.49	0.61	0.49	0.96	0.65
Combust_Cannabis	0.80	0.70	0.75	0.87	0.67	0.76	0.83	0.89	0.86
Combust_Tobacco	0.56	0.36	0.44	0.67	0.61	0.64	0.71	0.61	0.66
Smoking_Cessation	0.38	0.36	0.37	0.48	0.35	0.40	0.61	0.45	0.52
Brands	< 0.01	< 0.01	< 0.01	0.1	0.1	0.1	0.3	0.2	0.24
Overall	**0.75**	**0.37**	**0.49**	**0.89**	**0.61**	**0.72**	**0.89**	**0.68**	**0.77**

[1] PPV: positive prediction value (Precision)

[2] Sens. : sensitivity (Recall)

[3] F1: $2 \times \frac{Precision \times Recall}{Precision + Recall}$

Table 3: Performance evaluation of the rule-based, CRF, and LSTM-RNN algorithms when applied to the annotated corpus (1/3 held-out test set)

We applied the LSTM-RNN — the best performing classifier (by F-score) — to our entire Reddit dataset of 767,788 initiating posts in order to determine the distribution of concepts across the different subreddits. The resulting distribution is shown in **Table 4**, where it can be seen that, as would be expected, in tobacco subreddits users' discussion focuses on combustible tobacco use, whereas the cannabis subreddits, the primary focus of discourse is on combustible cannabis. For vaping device subreddits, vaping nicotine is the main focus, with cessation subreddit users discussing both vaping nicotine and combustible tobacco use. Product mentions were discovered in the majority of processed posts, with the proportion of posts containing concepts ranging from 63% of posts in the cannabis subreddits to 80% of posts in the smoking cessation subreddit[5].

[5] The trained model will be made publicly available via GitHub on acceptance of the paper

Annotation Concept	**Tobacco (%)**[a]	**Cannabis (%)**[b]	**Vaping (%)**[c]	**Cessation (%)**[d]
Vaping_Cannabis	8.5	2.5	7.3	<1
Vaping_Nicotine	14.3	7.3	66.8	10.2
Combustible_Cannabis	24.3	74.4	14.2	23.3
Combustible_Tobacco	50.4	13.5	6.5	48.2
Smoking_Cessation	1.8	<1	1.5	16.9
Brands	8.3	<1	2.7	<1

Annotation Concept	`cig` (%)	`mj` (%)	`trees` (%)	`weed` (%)	`ecig` (%)	`vape` (%)	`vape101` (%)	`stopsmoking` (%)[e]
Vaping_Cannabis	<1	<1	1.7	2.7	<1	3.1	4.2	<1
Vaping_Nicotine	27.8	2.5	10.4	6.6	72.6	80.6	69.6	9.6
Combustible_Cannabis	16.6	89.5	70.2	73.5	9.9	6.1	14.4	24.37
Combustible_Tobacco	47.9	6.2	13.2	14.1	12.2	5.9	7.3	49.78
Smoking_Cessation	<1	<1	<1	<1	1.5	1.3	1.7	14.3
Brands	5.8	<1	1.9	1.7	3.1	2.5	2.5	1.2

[a] The Tobacco category consists of data derived from the `r/stopsmoking` `r/cigarettes` subreddits.

[b] The Cannabis category consists of data derived from the `r/weed`, `r/trees`, and `r/marijuana` subreddits.

[c] The Vaping category consists of data derived from `r/vaping`, `r/electronic_cigarettes`, and `r/vaping101`.

[d] The Cessation category consists of data derived from `r/stopsmoking`

[e] `cig` refers to the `r/cigarettes` subreddit. `mj` refers to the `r/marijuana` subreddit. `trees` refers to the `r/trees` subreddit. `weed` refers to the `r/weed` subreddit. `ecig` refers to the `r/electronic_cigarettes` subreddit. `vape` refers to the `r/vaping` subreddit. `vape101` refers to the `r/vaping101` subreddit. `stopsmoking` refers to the `r/stopsmoking` subreddit.

Table 4: LSTM-RNN prediction results on 767,788 Reddit posts for the six annotation categories

3.4. Summary of Results

A number of interesting results emerged from our study. First, we observed that the LSTM-RNN classifier trained to identify product mentions resulted in a performance of 0.77 (F-score) compared to the CRF and rule-based baseline approaches (0.72 & 0.49 F-score, respectively, see **Table 3**). When the LSTM-RNN classifier was applied to our entire corpus of 767,788 initiating posts we found that combustible cannabis use was — as may be anticipated — extensively discussed in cannabis-related subreddits (74% of initiating posts). More surprisingly, combustible cannabis was also extensively discussed in the tobacco (24% of initiating posts) and smoking cessation (23% of initiating posts) subreddits (see **Table 4**).

Second, over 10% of posts in the smoking cessation subreddit discussed vaping nicotine, with <1% discussing vaping cannabis. This suggests that a substantial minority of users of the cessation subreddit are discussing ENDS products, although whether this discussion relates to current, former, or potential use is unclear. This result is consistent with survey research on the relationship between ENDS use and cessation (Weaver et al., 2020).

Third, we found that 17% of smoking cessation posts contain explicit statements regarding quitting, while only 2% of vaping-related posts discussed quitting. This result supports Chen et al. (2015)'s claim that the vast majority of posters in vaping subreddits are motivated to participate in a hobbyist community focused on sharing information, rather than seeking advice and support regarding smoking cessation.

	Example errors	Cause of error
1	. . . really enjoy my alien{VAPING_NIC/TOBACCO}{NO_LABEL}	Vocabulary unseen in training data
2	I've been using honey vape cartridge{VAPING_MJ}{VAPING_NIC/TOBACCO} for a few months now.	Vocabulary unseen in training data
3	. . . definitely can't go back to daily smoking{COMBUST_MJ}{COMBUST_TOBACCO}.	Text is from cannabis related forum.

Table 5: LSTM error analysis. Note that the first label in each example is derived from our manually annotated dataset, whereas the second label (rendered in red) was generated by the LSTM algorithm

Fourth, the performance of our algorithm was moderate rather than excellent, with some categories achieving relatively high F-scores (e.g. COMBUSTIBLE_CANNABIS & COMBUSTIBLE_TOBACCO) and others achieving relatively low F-scores (e.g. VAPING_CANNABIS & BRANDS). To better understand the strengths and limitations of the algorithm, we performed an error analysis (example errors show **Table 5**). In performing the error analysis, we noticed several distinct causes of error, including missed labelling of new products — especially new vaping products — not in the training data, and insufficient context to determine correct product type.

4. Discussion

This study presents work on applying an NLP algorithm designed to disambiguate between tobacco and cannabis product mentions (both smoking and vaping) in Reddit posts, hence contributing to our currently limited understanding of users' often complex product use patterns. Using an NLP algorithm, we were able to identify with reasonable reliability the extent to which product mentions were discussed across different product-related subreddits. Our work has demonstrated that, due to pervasive ambiguity, simple rule-based methods are inadequate for the task of disambiguating between vaping tobacco and vaping cannabis, and smoking tobacco and smoking cannabis. Instead, more sophisticated (but less readily intelligible) neural-network based classifiers were required to meet our goal.

We found our results to be broadly consistent with results gained from a multi-state probability-based sample designed to quantify cannabis use modalities under the auspices of the Behavioral Risk Factor Surveillance System (Schauer et al., 2020). This survey found that in a sample of 6,174 adult cannabis users, only 2.1% reported exclusively vaping cannabis, while 58.3% reported exclusively smoking cannabis. Our NLP-supported analysis of Reddit cannabis subreddits found that 2.5% of initiating posts contained product mentions referring to vaping cannabis and 74.4% of initiating posts contained product mentions referring to combustible cannabis, with the caveat that the sensitivity of our Vaping_Cannabis classifier was relatively low (0.21) suggesting that the actual proportion of posts mentioning vaping cannabis may be substantially higher than indicated by our results. Similarly, we found that 9.6% of posts in our smoking cessation dataset referred to vaping nicotine, a lower proportion than might be expected when compared to the prevalence of ENDS use among current smokers (27.7% as determined by Owusu et al. (2019)). It is possible that this discrepancy can be accounted for by the fact that individuals who post in `r/stopsmoking` are a subset of current (and recent) smokers who are sufficiently interested in quitting to contribute to a smoking cessation online community, and hence are less likely to engage in co- and dual-use behavior. These comparisons with results generated by surveys suggest that Reddit may have some utility as an ancillary tool for providing insights regarding changing product use patterns more rapidly and inexpensively than is possible through surveys based on high-quality probability samples (Amaya et al., 2019).

A possible explanation for the relative lack of cross-product discussion between the different subreddits is that users may be inhibited from discussing dual-use due to the fact that that some types of product use may be percieved as falling outside the purview of a particular subreddit, given the norms associated with that community. For example the `r/vaping` subreddit explicitly discourages posts related to dry herb vaporization and oil pen vaporization, directing posts related to cannabis vaporization products and practices to alternative subreddits. For any analysis using Reddit data, it is important to take into account these subreddit-specific contextual factors.

Key areas for future work — in addition to improving the performance of the algorithm — include using computational methods to determine how public sentiment towards particular products varies by subreddit (e.g. do users in the `r/stopsmoking` subreddit hold a broadly negative attitude towards combustible cannabis use?) and the application of similar product disambiguation algorithms to other sources of consumer-generated text (e.g. Twitter, online health communities).

The research described in this paper is not without limitations. When performing the annotation, it is likely that important product-related keywords were not included due to the rapidly changing terminology surrounding cannabis and vaping product usage. A further challenge that is associated with social media research generally is that Reddit users, skewing young and male (Pew, 2019), are unlikely to provide a representative sample of the general population. This potential lack of representativeness is exacerbated by the well-supported claim that internet forum participation is governed by the "90-9-1" principle (i.e. 90% of forum users "lurk" but do not actively participate, 9% of users participate sparingly, and only 1% of participants are highly active (van Mierlo, 2014)). Further, individual may not necessarily disclose the full range of their product use in a particular subreddit. For example, an individual may be a regular user of combustible cannabis and an occasional user of ENDS, but posts exclusively in the r/trees (cannabis) subreddit and not in vaping-related subreddits.

Despite its limitations, evidence presented in this study supports the contention that Reddit data — in conjunction with NLP — is a useful means of better understanding product discussion patterns and confirms the utility of Reddit as a data source for substance use research. Finally, this work goes some way towards establishing a baseline method for what is a key problem besetting social media tobacco control NLP: reliably distinguishing between cannabis and tobacco product use in social media text.

Acknowledgments

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.

References

Amaya, A., Bach, R., Keusch, F., Kreuter, F., 2019. New data sources in social science research: Things to know before working with Reddit data. Social Science Computer Review , 0894439319893305URL: https://doi.org/10.1177/0894439319893305, doi:10.1177/0894439319893305, arXiv:https://doi.org/10.1177/0894439319893305.

Ayers, J.W., Leas, E.C., Allem, J.P., Benton, A., Dredze, M., Althouse, B.M., Cruz, T.B., Unger, J.B., 2017. Why do people use Electronic Nicotine Delivery Systems (electronic cigarettes)? a content analysis of Twitter, 2012-2015. PLoS One 12, e0170702. doi:10.1371/journal.pone.0170702.

Barker, J.O., Rohde, J.A., 2019. Topic clustering of e-cigarette submissions among Reddit communities: A network perspective. Health Educ Behav 46, 59–68. doi:10.1177/1090198119863770.

Beletsky, L., Seymour, S., Kang, S., Siegel, Z., Sinha, M.S., Marino, R., Dave, A., Freifeld, C., 2020. Fentanyl panic goes viral: The spread of misinformation about overdose risk from casual contact with fentanyl in mainstream and social media. Int J Drug Policy 86, 102951. doi:10.1016/j.drugpo.2020.102951.

Benson, R., Hu, M., Chen, A.T., Nag, S., Zhu, S.H., Conway, M., 2020. Investigating the attitudes of adolescents and young adults towards JUUL: Computational study using Twitter data. JMIR Public Health Surveill 6, e19975. doi:10.2196/19975.

Benton, A., Coppersmith, G., Dredze, M., 2017. Ethical research protocols for social media health research, in: Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, Association for Computational Linguistics, Valencia, Spain. pp. 94–102. URL: https://www.aclweb.org/anthology/W17-1612, doi:10.18653/v1/W17-1612.

Bhatnagar, A., Payne, T.J., Robertson, R.M., 2019. Is there a role for electronic cigarettes in tobacco cessation? J Am Heart Assoc 8, e012742. doi:10.1161/JAHA.119.012742.

Breijyeh, Z., Jubeh, B., Bufo, S.A., Karaman, R., Scrano, L., 2021. Cannabis: A toxin-producing plant with potential therapeutic uses. Toxins (Basel) 13. doi:10.3390/toxins13020117.

Cavazos-Rehg, P., Grucza, R., Krauss, M.J., Smarsh, A., Anako, N., Kasson, E., Kaiser, N., Sansone, S., Winograd, R., Bierut, L.J., 2019. Utilizing social media to explore overdose and HIV/HCV risk behaviors among current opioid misusers. Drug Alcohol Depend 205, 107690. doi:10.1016/j.drugalcdep.2019.107690.

Chatzis, S.P., Demiris, Y., 2013. The infinite-order conditional random field model for sequential data modeling. IEEE Transactions on Pattern Analysis & Machine Intelligence 35, 1523–1534.

Chen, A.T., Zhu, S.H., Conway, M., 2015. What online communities can tell us about electronic cigarettes and hookah use: A study using text mining and visualization techniques. J Med Internet Res 17, e220. doi:10.2196/jmir.4517.

Conway, M., 2014. Ethical issues in using Twitter for public health surveillance and research: Developing a taxonomy of ethical concepts from the research literature. J Med Internet Res 16, e290.

Conway, M., Hu, M., Chapman, W.W., 2019. Recent advances in using natural language processing to address public health research questions using social media and consumer generated data. Yearb Med Inform 28, 208–217. doi:10.1055/s-0039-1677918.

Correa, D., Silva, L.A., Mondal, M., Benevenuto, F., Gummadi, K.P., 2015. The many shades of anonymity: Characterizing anonymous social media content., in: Cha, M., Mascolo, C., Sandvig, C. (Eds.), ICWSM, AAAI Press. pp. 71–80. URL: http://dblp.uni-trier.de/db/conf/icwsm/icwsm2015.html#CorreaSMBG15.

Czaplicki, L., Kostygina, G., Kim, Y., Perks, S.N., Szczypka, G., Emery, S.L., Vallone, D., Hair, E.C., 2019. Characterising JUUL-related posts on Instagram. Tob Control doi:10.1136/tobaccocontrol-2018-054824.

Delnevo, C.D., Bover-Manderski, M.T., Hrywna, M., 2011. Cigar, marijuana, and blunt use among us adolescents: Are we accurately estimating the prevalence of cigar smoking among youth? Prev Med 52, 475–6. doi:10.1016/j.ypmed.2011.03.014.

Doran, N., Myers, M.G., Correa, J., Strong, D.R., Tully, L., Pulvers, K., 2019. Marijuana use among young adult non-daily cigarette smokers over time. Addict Behav 95, 91–97. doi:10.1016/j.addbeh.2019.03.007.

Emery, S.L., Vera, L., Huang, J., Szczypka, G., 2014. Wanna know about vaping? patterns of message exposure, seeking and sharing information about e-cigarettes across media platforms. Tobacco Control 3, 17–25.

Foufi, V., Timakum, T., Gaudet-Blavignac, C., Lovis, C., Song, M., 2019. Mining of textual health information from Reddit: Analysis of chronic diseases with extracted entities and their relations. J Med Internet Res 21, e12876. doi:10.2196/12876.

Gaffney, D., Matias, J.N., 2018. Caveat emptor, computational social science: Large-scale missing data in a widely-published Reddit corpus. PloS One 13, e0200162.

Gkotsis, G., Oellrich, A., Velupillai, S., Liakata, M., Hubbard, T.J.P., Dobson, R.J.B., Dutta, R., 2017. Characterisation of mental health conditions in social media using informed deep learning. Sci Rep 7, 45141. doi:10.1038/srep45141.

Goffman, E., 1963. Stigma: notes on the management of spoiled identity. Prentice-Hall, Englewood Cliffs, N.J.

Grana, R., Benowitz, N., Glantz, S.A., 2014. E-cigarettes: a scientific review. Circulation 129, 1972–86. doi:10.1161/CIRCULATIONAHA.114.007667.

Grucza, R.A., Abbacchi, A.M., Przybeck, T.R., Gfroerer, J.C., 2007. Discrepancies in estimates of prevalence and correlates of substance use and disorders between two national surveys. Addiction 102, 623–9. doi:10.1111/j.1360-0443.2007.01745.x.

Hajek, P., Phillips-Waller, A., Przulj, D., Pesola, F., Myers Smith, K., Bisal, N., Li, J., Parrott, S., Sasieni, P., Dawkins, L., Ross, L., Goniewicz, M., Wu, Q., McRobbie, H.J., 2019. A randomized trial of e-cigarettes versus nicotine-replacement therapy. N Engl J Med 380, 629–637. doi:10.1056/NEJMoa1808779.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Computation 9, 1735–1780. URL: https://dl.acm.org/doi/10.1162/neco.1997.9.8.1735.

Hripcsak, G., Rothschild, A.S., 2005. Agreement, the f-measure, and reliability in information retrieval. J Am Med Inform Assoc 12, 296–8. doi:10.1197/jamia.M1733.

Huang, Z., Xu, W., Yu, K., 2015. Bidirectional LSTM-CRF models for sequence tagging. CoRR abs/1508.01991. URL: http://arxiv.org/abs/1508.01991.

Jurafsky, D., Martin, J.H., 2009. Speech and Language Processing: an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 2nd ed ed., Pearson Prentice Hall, Upper Saddle River, N.J. URL: http://www.loc.gov/catdir/toc/ecip0812/2008010335.html.

Kim, K., Gibson, L.A., Williams, S., Kim, Y., Binns, S., Emery, S.L., Hornik, R.C., 2020. Valence of media coverage about electronic cigarettes and other tobacco products from 2014-2017: Evidence from automated content analysis. Nicotine Tob Res doi:10.1093/ntr/ntaa090.

Knotta, C.E., Gomori, S., Ngyuen, M., Pedrazzani, S., Sattaluri, S., Mierzwa, F., Chantala, K., 2021. Connecting and linking neurocognitive, digital phenotyping, physiologic, psychophysical, neuroimaging, genomic, & sensor data with survey data. Social Science Computer Review .

Lafferty, J., McCallum, A., Pereira, F.C., 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in: ICML '01: Proceedings of the Eighteenth International Conference on Machine

Learning, pp. 282—289. URL: https://arxiv.org/pdf/1709.03637.pdf.

MacLean, D., Gupta, S., Lembke, A., Manning, C., Heer, J., 2015. Forum77: An analysis of an online health forum dedicated to addiction recovery, in: CSCW '15: Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work and Social Computing, pp. 1511–1526. URL: https://jan.stanford.edu/pubs/2015-Forum77-CSCW.pdf.

McDonald, E.A., Popova, L., Ling, P.M., 2016. Traversing the triangulum: the intersection of tobacco, legalised marijuana and electronic vaporisers in Denver, Colorado. Tob Control 25, i96–i102. doi:10.1136/tobaccocontrol-2016-053091.

Meacham, M.C., Paul, M.J., 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 Alcohol Depend 188, 364–369. doi:10.1016/j.drugalcdep.2018.03.041.

van Mierlo, T., 2014. The 1% rule in four digital health social networks: an observational study. J Med Internet Res 16, e33. doi:10.2196/jmir.2966.

Myslín, M., Zhu, S.H., Chapman, W., Conway, M., 2013. Using Twitter to examine smoking behavior and perceptions of emerging tobacco products. J Med Internet Res 15, e174. doi:10.2196/jmir.2534.

Nguyen, T., Larsen, M.E., O'Dea, B., Phung, D., Venkatesh, S., Christensen, H., 2017. Estimation of the prevalence of adverse drug reactions from social media. Int J Med Inform 102, 130–137. doi:10.1016/j.ijmedinf.2017.03.013.

Okazaki, N., 2016. CRFsuite a fast implementation of conditional random fields (CRFs). URL: http://www.chokkan.org/software/index.html.en.

Oram, D., Tzilos Wernette, G., Nichols, L.P., Vydiswaran, V.G.V., Zhao, X., Chang, T., 2018. Substance use among young mothers: An analysis of Facebook posts. JMIR Pediatr Parent 1, e10261. doi:10.2196/10261.

Owusu, D., Huang, J., Weaver, S.R., Pechacek, T.F., Ashley, D.L., Nayak, P., Eriksen, M.P., 2019. Patterns and trends of dual use of e-cigarettes and cigarettes among U.S. adults, 2015-2018. Prev Med Rep 16, 101009. doi:10.1016/j.pmedr.2019.101009.

Park, A., Conway, M., 2017. Longitudinal changes in psychological states in online health community members: Understanding the long-term effects of participating in an online depression community. J Med Internet Res 19, e71. doi:10.2196/jmir.6826.

Park, A., Hartzler, A.L., Huh, J., Hsieh, G., McDonald, D.W., Pratt, W., 2016. "how did we get here?": Topic drift in online health discussions. J Med Internet Res 18, e284. doi:10.2196/jmir.6297.

Paul, M., Dredze, M., 2018. Social Monitoring for Public Health. Morgan & Claypool Publishers. URL: https://www.cs.jhu.edu/~mdredze/social-monitoring-for-public-health/.

Pew, 2019. Who uses YouTube, WhatsApp and Reddit. https://www.pewresearch.org/internet/chart/who-uses-youtube-whatsapp-and-reddit/.

Pushshift, 2021. Pushshift.io api. URL: https://github.com/pushshift/api.

SAMHSA, 2015. Behavioral Health Trends in the United States: Results from the 2014 National Survey on Drug Use and Health. Technical Report. Department of Health and Human Services. URL: https://www.samhsa.gov/data/sites/default/files/NSDUH-FRR1-2014/NSDUH-FRR1-2014.pdf.

Sarker, A., O'Connor, K., Ginn, R., Scotch, M., Smith, K., Malone, D., Gonzalez, G., 2016. Social media mining for toxicovigilance: Automatic monitoring of prescription medication abuse from Twitter. Drug Saf 39, 231–40. doi:10.1007/s40264-015-0379-4.

Schauer, G.L., Njai, R., Grant-Lenzy, A.M., 2020. Modes of marijuana use – smoking, vaping, eating, and dabbing: Results from the 2016 BRFSS in 12 states. Drug and Alcohol Dependence 209.

Smith, D.M., Miller, C., O'Connor, R.J., Kozlowski, L.T., Wadsworth, E., Fix, B.V., Collins, R.L., Wei, B., Goniewicz, M.L., Hyland, A.J., Hammond, D., 2020. Modes of delivery in concurrent nicotine and cannabis use ("co-use") among youth: Findings from the International Tobacco Control (ITC) Survey. Subst Abus , 1–9doi:10.1080/08897077.2019.1709603.

Sowles, S.J., McLeary, M., Optican, A., Cahn, E., Krauss, M.J., Fitzsimmons-Craft, E.E., Wilfley, D.E., Cavazos-Rehg, P.A., 2018. A content analysis of an online pro-eating disorder community on reddit. Body Image 24, 137–144. doi:10.1016/j.bodyim.2018.01.001.

Surgeon General, 2014. The Health Consequences of Smoking – 50 years of progress. Technical Report. U.S. Department of Health and Human Services, United States Public Health Service Office of the Surgeon General. URL: https://www.hhs.gov/surgeongeneral/reports-and-publications/tobacco/index.html.

Surgeon General, 2020. Smoking Cessation: A Report of the Surgeon General. Technical Report. U.S. Department of Health and Human Services, United States Public Health Service Office of the Surgeon General.

Tamersoy, A., De Choudhury, M., Chau, P., 2015. Characterizing smoking and drinking abstinence from social media, in: Proceedings of 26th ACM Conference on Hypertext and Social Media, pp. 139–148.

Tankovska, H., 2021. Regional distribution of desktop traffic to reddit.com as of december 2020, by country. URL: https://www.statista.com/statistics/325144/reddit-global-active-user-distribution/.

TensorFlow, 2019. Tensorflow. URL: https://www.tensorflow.org/.

Tofighi, B., El Shahawy, O., Segoshi, A., Moreno, K.P., Badiei, B., Sarker, A., Krawczyk, N., 2021. Assessing perceptions about medications for opioid use disorder and Naloxone on Twitter. J Addict Dis 39, 37–45. doi:10.1080/10550887.2020.1811456.

Van Swol, L.M., Chang, C.T., Kerr, B., Moreno, M., 2020. Linguistic predictors of problematic drinking in alcohol-related Facebook posts. Journal of Health Communication 25, 214–222.

Volkow, N.D., Baler, R.D., Compton, W.M., Weiss, S.R.B., 2014. Adverse health effects of marijuana use. N Engl J Med 370, 2219–27. doi:10.1056/NEJMra1402309.

Wadsworth, E., Neale, J., McNeill, A., Hitchman, S.C., 2016. How and why do smokers start using e-cigarettes? qualitative study of vapers in London, UK. Int J Environ Res Public Health 13. doi:10.3390/ijerph13070661.

Wang, X., Zhao, K., Cha, S., Amato, M.S., Cohn, A.M., Pearson, J.L., Papandonatos, G.D., Graham, A.L., 2019. Mining user-generated content in an online smoking cessation community to identify smoking status: A machine learning approach. Decision Support Systems 116, 26–34.

Warner, K.E., 2014. Tobacco control policies and their impacts. past, present, and future. Ann Am Thorac Soc 11, 227–30. doi:10.1513/AnnalsATS.201307-244PS.

Weaver, S.R., Heath, J.W., Ashley, D.L., Huang, J., Pechacek, T.F., Eriksen, M.P., 2020. What are the reasons that smokers reject ends? a national probability survey of u.s. adult smokers, 2017-2018. Drug Alcohol Depend 211, 107855. doi:10.1016/j.drugalcdep.2020.107855.

Wongkoblap, A., Vadillo, M.A., Curcin, V., 2017. Researching mental health disorders in the era of social media: Systematic review. J Med Internet Res 19, e228. doi:10.2196/jmir.7215.

Yang, X., Bian, J., Gong, Y., Hogan, W.R., Wu, Y., 2019. Madex: A system for detecting medications, adverse drug events, and their relations from clinical notes. Drug Safety 42, 123–133.