Studying Anonymous Health Issues and Substance Use on College Campuses with Yik Yak

Animesh Koratana[1], **Mark Dredze**[1], **Margaret S. Chisolm**[2], **Matthew W. Johnson**[2], **Michael J. Paul**[3*]
[1] Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218
[2] Department of Psychiatry and Behavioral Sciences, Johns Hopkins University, Baltimore, MD 21224
[3] Department of Information Science, University of Colorado, Boulder, CO 80309
[*] Corresponding author: mpaul@colorado.edu

Abstract

This study investigates the public health intelligence utility of Yik Yak, a social media platform that allows users to anonymously post and view messages within precise geographic locations. Our dataset contains 122,179 "yaks" collected from 120 college campuses across the United States during 2015. We first present an exploratory analysis of the topics commonly discussed in Yik Yak, clarifying the health issues for which this may serve as a source of information. We then present an in-depth content analysis of data describing substance use, an important public health issue that is not often discussed in public social media, but commonly discussed on Yik Yak under the cloak of anonymity.

Introduction

Social media data have been demonstrated to be viable sources of intelligence for a variety of public health issues, such as disease surveillance (Paul, Dredze, and Broniatowski 2014) and medication safety monitoring (O'Connor et al. 2014). However, with popular platforms such as Twitter, health intelligence is limited by what people are willing to share publicly. Stigmatizing health conditions, like sexually transmitted disease, are less likely to be shared in public spaces (De Choudhury, Morris, and White 2014), reducing the utility of social media for these important issues.

In this study, we consider a social media platform that provides a fully anonymous experience: **Yik Yak**, which has been studied only recently as a source of trending information (Leskovich 2015; McKenzie, Adams, and Janowicz 2015). While many social media platforms offer pseudonymity (in which people use aliases as identifiers rather than their real names), few offer full anonymity. Our hypothesis is that an anonymous but publicly viewable platform such as Yik Yak may provide unique insights into health issues that are not commonly discussed elsewhere.

Moreover, Yik Yak has a unique system of location-centered content delivery, in which messages are only viewable to users within a 5-mile radius of where the message was posted. This enables us to analyze discussions happening within precise locations. As Yik Yak is popular among college students, we chose to focus our study on data within close proximity to dozens of college campuses in the United States, giving insights into the health issues discussed across campuses. We focus in depth on an important but stigmatizing public health problem: substance use, both licit (e.g., alcohol, tobacco) and illicit.

This study consists two parts. First, we broadly explore the range of health issues that are commonly discussed in our college-centered Yik Yak dataset (described in the next section). Using a topic modeling approach, we reveal key themes of content and identify which health issues are and are not present in the data. We then focus on the particular issue of substance use, and present a content analysis of substance-related data.

Data Collection

Our data come from Yik Yak, a social media application that allows users to post and view messages (called "yaks") within a 5-mile radius of the message location. This location-focused approach encourages discussions within physical communities, and for the purpose of this study, this property also enables us to collect yaks from precise geographic locations. Another important property of Yik Yak is that the messages are posted without any user identifier, creating a fully anonymous experience.

While yaks are typically delivered based on the location of the user, our crawler queried for yaks with certain parameters, in which we could specify the latitude and longitude of the center of the 5-mile yak radius. This allowed us to obtain yaks from multiple locations. We collected yaks from **120 college campuses** in the United States. This set of campuses includes the largest universities in the US, along with additional universities that we added to increase the breadth and diversity of our collection. For each campus, we queried for yaks within the radius of the campus' geo-coordinates, which we obtained from Google Maps Geocoding API.

We continuously crawled yaks from June 12, 2015 to July 14, 2015. The crawler returns the 100 most recent yaks for a given query. Through this process, we obtained **122,179 total yaks**. In addition to the original yak, users can also reply to yaks, forming comment chains. We collected both original yaks and their replies (the counts do not include replies). We note that our study analyses exactly the data available from Yik Yak, which maintains the anonymity of the service.

Health Topics								
Topic 9 "SLEEP"	**Topic 16** "HYGIENE"	**Topic 26** "SEX"	**Topic 27** "DRUGS"	**Topic 34** "SEX"	**Topic 36** "FOOD"	**Topic 39** "DRINK"	**Topic 40** "WEIGHT"	**Topic 48** "HYGIENE"
sleep	poop	sex	weed	dick	eat	drink	fat	smell
day	shit	like	smoke	suck	food	drunk	weight	use
bed	toilet	girl	drugs	mouth	pizza	coffee	gym	like
night	bathroom	get	smoking	kiss	good	beer	eat	shower
im	ass	girls	drug	tip	eating	drinking	body	water
work	like	guys	doctor	head	cheese	water	lose	teeth
go	shower	guy	high	give	chicken	alcohol	healthy	wash
morning	eat	time	take	pussy	chipotle	wine	eating	skin
home	paper	want	anxiety	blow	like	milk	im	hair
time	pee	feel	got	want	want	starbucks	workout	face
Other Topics								
Topic 1 "PARTYING"	**Topic 13** "SOCIAL"	**Topic 17** "FAMILY"	**Topic 18** "SEXUALITY"	**Topic 19** "RACE"	**Topic 23** "HOUSING"	**Topic 38** "RELIGION"	**Topic 42** "JOBS"	**Topic 44** "SCHOOL"
go	friends	mom	gay	white	room	god	money	class
party	friend	dad	straight	black	live	church	job	major
tonight	people	parents	im	people	roommate	religion	get	classes
going	best	family	guy	racist	house	jesus	pay	summer
good	im	kids	guys	race	apartment	marriage	work	school
bar	like	baby	girl	asian	living	believe	buy	take
fun	feel	child	bi	im	roommates	religious	much	taking
place	new	sister	want	racism	home	bible	make	study
night	make	kid	lesbian	privilege	place	christian	need	hard
bars	lonely	brother	dude	color	move	gay	would	professor

Table 1: Example topics learned by Latent Dirichlet Allocation on our Yik Yak corpus. These examples include the top 10 words of all 9 health topics and a sample of 9 other topics, with manually assigned topic labels.

Health Topics in Yik Yak

We first explore the content of the dataset using a topic modeling approach to identify prominent themes in the yaks. Topic models have previously been applied to social media data to understand which health issues are prominently discussed (Paul and Dredze 2011; 2014; Prier et al. 2011; Ghosh and Guha 2013; Wang, Paul, and Dredze 2014). The goal is to characterize the topic content of the corpus so that we understand the set of public health problems for which Yik Yak is a potential source of information.

We used Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003), a widely-used probabilistic topic model in which each text document has a distribution over latent "topics" and each topic has a distribution over words. By inferring the parameters of the model, we can learn which words cluster together as topics, and which topics are present in each document. We applied LDA to our corpus using 50 topics, treating each yak concatenated with its replies as a single document. We examined the top words for each topic and selected topics which were potentially pertinent to public health. Out of 50 topics, 9 topics were identified as being potentially relevant to health.[1]

[1]We attempted to identify more health topics by running LDA on a subset of the corpus that had been filtered for health, following the approach of Paul and Dredze (2014) which filters messages for health-related terms. However, this approach substantially reduced the amount of data without resulting in a broader set of health topics upon inspection. It also appears that there is simply not a high diversity of health issues commonly discussed in Yik Yak.

The health topics we identified covered the following public health problems: sleep, personal hygiene (2 topics), sexual activity (2 topics), substance use, food consumption, beverage consumption (including alcohol), and body weight/image. While not all of these topics are explicitly about health, some of them would be of interest in certain areas of public health. For example, the topic about food consumption would be relevant for the study of dietary patterns. Other topics not identified as health include recreation, relationships, and politics. Examples of both health-related topics and other topics are shown in Table 1. We note with interest that so many of the topics were relevant to health.

We make two observations about these topics. First, there is a high proportion of topics on sensitive subjects such as sex, substance use, and bathroom habits. This supports our hypothesis that anonymity supports disclosure of such subjects, although the focus on college campuses may also skew the topic content. Second, there is an absence of health topics that are commonly observed in other social media platforms like Twitter, such as upper respiratory infections and influenza-like illness. These findings suggest that Yik Yak is a poor source of general health intelligence, but a potential source for understanding certain high-stigma issues.

Substance Use Disclosure in Yaks

We now turn our attention to analyzing patterns in substance use, as disclosed through yaks. This is a public health issue for which social media holds promise, because traditional methods for monitoring substance use are often years out of date (Dunn et al. 2011). However, substance use is

Code	Substance	Yak
Use, neutral	Alcohol	Who else is already several beers deep?
Use, positive	Marijuana	I love smoking bud. It's the rare time when I'm not physically or emotionally hurting. I'll take a hit, forget why I'm depressed. I can actually smile again.
Use, negative	LSD	Did acid and think it might've messed with me. Would not recommend it for the feeble minded
Solicitation	Alcohol	I would like someone to buy me booze
Social group	Marijuana	Anyone want to smoke? I got the weed and now I want some company
Addiction	Tobacco	It's a funny realization when you realize that you'll just never quit smoking cigarettes
Info-seeking	Marijuana	What is it like to be high? I've never smoked pot.

Table 2: Representative examples (paraphrased to preserve privacy) of yaks coded with various categories.

less commonly discussed in popular social media, especially illegal drugs, though there are exceptions (Cavazos-Rehg et al. 2015). Previous work has primarily focused on online forums for substance users (Deluca et al. 2012; Paul and Dredze 2013), but forums do not provide precise location metadata and thus cannot focus on college locations as we do here. Moreover, due to lack of full anonymity (among other reasons), most forums prohibit the discussion of buying and selling of illicit drugs, which is an issue we are able to examine in this study.

While the topic model results provide an overview of general content, the results are not granular enough to isolate specific substances—for instance, "starbucks" and "alcohol" appear in the same topic. For more accurate content analysis, we used a manual coding approach. To find yaks relevant to substances, we filtered our corpus using a large set of substance-related keywords, and then further reduced the size of this collection by manually reading through all yaks and noting those related to substance use. For this study, substances include alcohol and tobacco as well as marijuana and other illicit drugs. This process resulted in **2,047 substance-related yaks**. Two annotators (one annotator per yak) then coded a sample of 500 of the substance yaks for categories of interest (one code per yak):

- **Substance use:** if the user discloses that (s)he has used a substance, past or present. If so, we also record whether the user describes the substance use as a positive or negative experience, if either.
- **Solicitation:** if the user seeks to acquire or provide a substance.
- **Social groups:** if the user is looking for others with whom to use a substance with.
- **Addiction:** if a yak references a substance addiction or an attempt to discontinue use, past or present.
- **Information-seeking:** if the user asks for information about a substance, such as legal or safety information.

For yaks that fit multiple categories, we selected the most salient code; yaks that fit none of these categories are simply labeled "Other". In addition to these categories, we noted which drug(s) are referenced in the yak, if any.

Of the 500 annotated yaks, **234 yaks** matched one of these categories, which we now summarize. Example yaks are shown in Table 2, illustrating the various content categories.

	Alcohol	Tobacco	Marijuana	Other
Code Distribution				
Use	50.6%	30.0%	44.4%	55.0%
Solicitation	17.3%	3.3%	23.1%	20.0%
Social groups	17.3%	23.3%	14.5%	5.0%
Addiction	6.2%	26.7%	2.6%	10.0%
Info-seeking	8.6%	16.7%	15.4%	10.0%
N	81	30	117	20
Use: Positive or Negative Experiences				
Positive	4.9%	0.0%	7.7%	9.1%
Negative	14.6%	0.0%	5.8%	18.2%
Neutral	80.5%	100.0%	86.5%	72.7%
N	41	9	52	11

Table 3: The distribution of codes for substance-related yaks, by substance type. N is the number of annotated yaks per drug (14 yaks had multiple drugs). Bottom: the distribution of sentiment codes for the subset of yaks labeled as describing substance use.

Substance Frequencies

The most commonly mentioned substances are marijuana (117 yaks), alcohol (81 yaks), and tobacco-derived products (30 yaks, including 2 yaks specifically mentioning electronic cigarettes). Other substances include Adderall (8 yaks), LSD (6 yaks), psilocybin (2 yaks), methamphetamine (1 yak), flakka (1 yak) and sleeping pills (1 yak).

Polysubstance use An important issue in public health is which substances are taken in combination. To explore this property, we counted the number of yaks which mentioned two or more substances. The most common pairs of substances are alcohol/marijuana (7 yaks), alcohol/tobacco (2 yaks), and marijuana/tobacco (2 yaks), while there were also mentions of alcohol/Adderall, alcohol/sleeping pills, marijuana/sleeping pills, and marijuana/psilocybin (1 yak each).

Code Breakdown

The distribution of codes among yaks labeled with different substance types are shown in Table 3. We observe some variation in the distribution across different types of substances.

Notably, tobacco had a far higher proportion of addiction-related yaks than other substances. Most of these yaks described attempts to quit smoking, or were celebrating abstinence from smoking ("smoke free, one year later!"). While we did not quantify the replies to yaks, we observed they were often supportive ("stay strong!").

We noted that the vast majority of solicitation yaks were for acquisition (e.g., buying) rather than providing (e.g., selling). Compared to the other substances, this type of yak was extremely rare for tobacco. This is likely because there are fewer legal barriers among a college-age cohort to obtaining tobacco (minimum age 18) than alcohol (minimum age 21) and illegal substances.

We noted a number of people use Yik Yak to find friends to use substances with ("I don't like smoking alone"). This type of yak was less common for the "other" type than for the three most popular substances, perhaps because the illicit nature of many other drugs makes this a riskier proposition.

Alcohol had the smallest proportion of information-seeking yaks, perhaps because this is is one of the most widely used substances among college students and fewer questions need to be asked. A very broad range of yaks were coded with this category: opinions on substances, clarifications of law, recommendations for places to use drugs, recommendations for which drug to use, etc. We did not code the replies to yaks in this analysis, though replies might be an additional source for analyzing opinions on substances.

Sentiment The vast majority of yaks that referenced substance use did not express a positive or negative experience, but of those that did, the sentiment was usually negative. Marijuana was an exception, with slightly more positive.

Discussion and Conclusion

We have presented what we believe to be the first study of Yik Yak as a potential source of public health intelligence, investigating this question broadly, through exploratory data analysis, and in depth, for the specific public health issue of substance use, finding high rates of substance use disclosure. Our analysis illuminated differences between different substances, such as a high proportion of addiction-related messages for tobacco, and a high proportion of solicitation-related messages for marijuana. Our findings suggest that Yik Yak is a potential source of candid information for studying substance use or other stigmatizing health issues.

While yaks have properties that make them unique from other social media, they also pose a number of limitations. Demographic attributes, which are important for epidemiological research, are unknown due to anonymity. While there is research on inferring such attributes from text (Rao et al. 2011; Culotta, Kumar, and Cutler 2015; Volkova et al. 2015) this often requires longer posting histories from users than isolated messages, yet with Yik Yak, there is generally no way to know if different messages are posted by the same person (unless users choose to provide identifiers through messages (Bernstein et al. 2011)). Similarly, while we targeted college campuses, we do not know which users are students. Our focus on campuses was intentional, but this means that it is not clear to what extent our topic and content analysis generalizes to Yik Yak as a whole. Because yaks are delivered around specified geographic points, it is a challenge to collect data that is geographically representative.

Despite these challenges, we believe that platforms with similar properties as Yik Yak, especially full anonymity, can provide an outlet for the sharing of important information for understanding population health.

References

Bernstein, M. S.; Monroy-Hernández, A.; Harry, D.; André, P.; Panovich, K.; and Vargas, G. G. 2011. 4chan and /b/: An analysis of anonymity and ephemerality in a large online community. In *ICWSM*.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research* 3:993–1022.

Cavazos-Rehg, P. A.; Krauss, M.; Fisher, S. L.; Salyer, P.; Grucza, R. A.; and Bierut, L. J. 2015. Twitter chatter about marijuana. *J Adolesc Health* 56(2):139–145.

Culotta, A.; Kumar, N. R.; and Cutler, J. 2015. Predicting the Demographics of Twitter Users from Website Traffic Data. In *AAAI*.

De Choudhury, M.; Morris, M. R.; and White, R. W. 2014. Seeking and sharing health information online: Comparing search engines and social media. In *CHI*, 1365–1376.

Deluca, P.; Davey, Z.; Corazza, O.; Di Furia, L.; Farre, M.; Flesland, L. H.; and *et al.* 2012. Identifying emerging trends in recreational drug use; outcomes from the Psychonaut Web Mapping Project. *Prog. Neuropsychopharmacol. Biol. Psychiatry* 39(2):221–226.

Dunn, M.; Bruno, R.; Burns, L.; and Roxburgh, A. 2011. Effectiveness of and challenges faced by surveillance systems. *Drug Testing and Analysis* 3(9):635–641.

Ghosh, D. D., and Guha, R. 2013. What are we 'tweeting' about obesity? Mapping tweets with Topic Modeling and Geographic Information System. *Cartogr Geogr Inf Sci* 40(2):90–102.

Leskovich, W. R. 2015. Yik Yak: a social media sensor. In *SPIE*, 9499.

McKenzie, G.; Adams, B.; and Janowicz, K. 2015. Of oxen and birds: Is Yik Yak a useful new data source in the geosocial zoo or just another Twitter? In *ACM SIGSPATIAL International Workshop on Location-Based Social Networks*.

O'Connor, K.; Pimpalkhute, P.; Nikfarjam, A.; Ginn, R.; Smith, K. L.; and Gonzalez, G. 2014. Pharmacovigilance on Twitter? Mining tweets for adverse drug reactions. *AMIA* 2014.

Paul, M. J., and Dredze, M. 2011. You are what you Tweet: Analyzing Twitter for public health. In *ICWSM*.

Paul, M., and Dredze, M. 2013. Drug extraction from the web: Summarizing drug experiences with multi-dimensional topic models. In *NAACL*.

Paul, M. J., and Dredze, M. 2014. Discovering health topics in social media using topic models. *PLoS ONE* 9(8):e103408.

Paul, M. J.; Dredze, M.; and Broniatowski, D. 2014. Twitter improves influenza forecasting. *PLOS Currents Outbreaks*.

Prier, K. W.; Smith, M. S.; Giraud-Carrier, C.; and Hanson, C. L. 2011. Identifying health-related topics on Twitter: An exploration of tobacco-related tweets as a test topic. In *SBP*, 18–25.

Rao, D.; Paul, M.; Fink, C.; Yarowsky, D.; Oates, T.; and Coppersmith, G. 2011. Hierarchical Bayesian models for latent attribute detection in social media. In *ICWSM*.

Volkova, S.; Bachrach, Y.; Armstrong, M.; and Sharma, V. 2015. Inferring latent user properties from texts published in social media. In *AAAI*.

Wang, S.; Paul, M. J.; and Dredze, M. 2014. Exploring health topics in Chinese social media: An analysis of Sina Weibo. In *AAAI Workshop on the World Wide Web and Public Health Intelligence*.