Predicting Public Opinion on Drug Legalization: Social Media Analysis and Consumption Trends

Farahnaz Golrooy Motlagh*, Saeedeh Shekarpour†, Amit Sheth*, Krishnaprasad Thirunarayan*, Michael L. Raymer*

* Wright State University, USA, Kno.e.sis Center, {farahnaz, amit, tkprasad, micheal}@knoesis.org

† University of Dayton, USA, sshekarpour1@udayton.edu

Abstract—In this paper, we focus on the collection and analysis of relevant Twitter data on a state-by-state basis for (i) measuring public opinion on marijuana legalization by mining sentiment in Twitter data and (ii) determining the usage trends for six distinct types of marijuana. We overcome the challenges posed by the informal and ungrammatical nature of tweets to analyze a corpus of 306,835 relevant tweets collected over the four-month period, preceding the November 2015 Ohio Marijuana Legalization ballot and the four months after the election for all states in the US. Our analysis revealed two key insights: (i) the people in states that have legalized recreational marijuana express greater positive sentiments about marijuana than the people in states that have either legalized medicinal marijuana or have not legalized marijuana at all; (ii) the states that have a high percentage of positive sentiment about marijuana is more inclined to authorize (e.g., by allowing medical marijuana) or broaden its legal usage (e.g., by allowing recreational marijuana in addition to medical marijuana). Our analysis shows that social media can provide reliable information and can serve as an alternative to traditional polling of public opinion on drug use and epidemiology research.

Index Terms—Marijuana Legalization, Drug Abuse Ontology, Public Opinion, Sentiment Analysis, Prediction, Consumption Trends, Entity Extraction, and Machine Learning

I. INTRODUCTION

There has been a surge of public support for changing the legal status of various forms of marijuana in the US, which has resulted in full or partial legalization in 25 states¹.

Marijuana has both recreational and medical uses. It is one of the most commonly used psychoactive substance; hence, there is heightened concern about its increased usage. However, marijuana has also gained acceptance for medicinal use as a pain reliever for severe diseases such as cancer, nerve pain, and glaucoma [11], [23]. Marijuana's possible medical

benefits are one reason for the surge of support for loosening marijuana restrictions throughout the United States.

Currently, 21 states² have legalized marijuana exclusively for medical purposes and recreational use as well, including California, Arizona, Illinois, and Nevada, and four other states (Alaska, Colorado, Oregon, and Washington) [17]. In this study, we discuss three legalization efforts for recreational marijuana (Section 2).

In Alaska, adults older than 21 are allowed to purchase and consume up to one ounce of marijuana. In Oregon, voting resulted in the legalization of up to eight ounces for consumption only inside the home. Other states had passed laws which allow the limited use of cannabis oil (a specific type of marijuana) when people diagnosed with particular illnesses. Yet we do not have adequate insight into macro-level factors which influence the community norms that directly affect legalization considerations. In this work, we mine social media to reveal public opinion and provide a useful alternative to traditional surveys and polling. This effort is part of the NIDA-sponsored eDrugTrends project³, which focuses on public health surveillance and monitoring by collecting tweets related to marijuana and its usage. We rely on Twitter as our underlying social media data source because of its growing use for public health surveillance [1], [2], [10]. Currently, in the US Twitter is used by approximately 24% of the population⁴. Twitter reports 313 million monthly active users that generate over 500 million tweets per day⁵. Although the size of a tweet is restricted to 140 characters, meaning that each tweet can provide only a limited amount of information, once we consider a more substantial aggregation of tweets on a particular topic, mining approaches can uncover valuable community-level insights.

The focus of our current study is to predict public opinion about the subject of *marijuana legalization* on a state-by-state basis by measuring the sentiment expressed in tweets and to measure the popularity and trends in the use of various types of marijuana at the state level.

¹https://en.wikipedia.org/wiki/Timeline_of_cannabis_laws_in_the_United_States

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ASONAM '19, August 27-30, 2019, Vancouver, Canada

© 2019 Association for Computing Machinery.

ACM ISBN 978-1-4503-6868-1/19/08?/\$15.00

<https://doi.org/10.1145/3341161.3344380>

²https://en.wikipedia.org/wiki/Timeline_of_cannabis_laws_in_the_United_States

³<http://wiki.knoesis.org/index.php/EDrugTrends/>

⁴<https://www.pewinternet.org/2016/11/11/social-media-update-2016/>

⁵<https://about.twitter.com/company>

To efficiently crawl the appropriate data, we created a comprehensive lexicon containing all the common terms, including slang and colloquialisms, which are used when referring to marijuana. Furthermore, we identified six distinct types of marijuana, which we derived from an underlying ontology developed for the eDrugTrends project. Our filtering approach relies on both a lexicon-based and a semantics-based approach. We mine public opinion by employing a coherent supervised classifier to predict the polarity (i.e., positive, negative, and neutral sentiments) of the tweets (Section V-A).

Some of the insights uncovered through our analysis include: (1) the people in states which have legalized recreational marijuana display greater happiness (i.e., positive sentiments) than the people from the other states; and (2) the states which have the highest percentage of positive sentiment for marijuana are more inclined to legalize or broaden its usage (section V). For example, that if a state displays a high degree of positive sentiment and has only legalized medical marijuana, then it is likely that it will extend legalization to cover recreational marijuana as well; likewise, if a state has not sanctioned any form of medical marijuana, but there is a high degree of positive sentiment, then its voters are likely to favor at least legalizing medical marijuana.

To the best of our knowledge, this is the first work in the biomedical domain which integrates lexicon with structured representation (i.e., Drug Abuse Ontology) for tasks such as (i) creating comprehensive seed terms, and (ii) entity identification on the textual corpus. This ontology-based retrieval approach aided us to reach higher recall. Although prior work [18] utilizes DBpedia⁶ (which is an structured version of Wikipedia) for information extraction, which is not conducive for our experiment. This is because, it does not include informal terms(e.g. slang terms) which play a substantial role in our context [12].

This paper is structured as follows. Section II describes our initial goals for the study and the corresponding challenges. Section III provides an overview of the architecture. Section IV describes the steps involved in data collection. Section V elaborates on the details of our approach, analysis, and the insights derived from our experimental study. Section VI reviews the related work. Section VII provides conclusion and a discussion of future work.

II. AIMS AND CHALLENGES

With respect to marijuana legalization, states in the US can be grouped into three categories: (i) states where marijuana is still illegal both for medical as well as recreational purposes, (ii) states where marijuana has been legalized only for medical purposes, and (iii) states where marijuana has been legalized for both medical and recreational use. Since social media reflects public opinion, it can be used to analyze voter trends with respect to marijuana legalization [16]. Therefore, we base our processing, analysis, and prediction of social media data and utilize it as an alternative to polling for measuring

public sentiment regarding marijuana legalization. This kind of monitoring of public opinion before and after an election can further help us to predict future changes in legalization. For example, we would expect to observe a noticeable change in the public opinion on the issue of marijuana legalization that is influenced by polling results on election day. We are also interested in investigating whether people residing in states where both recreational and medicinal marijuana have been legalized continue to express support for the drug's legalization (i.e., positive sentiment) as compared to people residing in states where marijuana has not yet been authorized. Moreover, measuring public opinion on marijuana can provide insights into whether a state will be motivated to pursue the legalization of recreational and/or medicinal marijuana in the future. Additionally, we are interested in the trending popularity of various types of marijuana. The forces behind our research are back up for significant trends on social media as marijuana legalization a vote in various jurisdictions. The question that arises is how can we support our claim about ranking states based on the growing shift towards marijuana legalization and the category of marijuana that could potentially be legalized based on the opinion expressed on social media across states. This type of analysis reveals insights that are of significant interest to social scientists and policymakers in the United States(US). In this study, we encountered substantial challenges in collecting a comprehensive set of tweets related to marijuana due to the unstructured nature of tweets, their frequent use of slang terms, and the brevity of tweets.

Fig. 1: Overview of our experiment.

III. ARCHITECTURE

Our proposed system contains the three main modules illustrated in Figure 1.

- 1) *Data Collection*: This module collects relevant and context-specific tweets which deal with the subject of marijuana legalization. It filters tweets based on three criteria: (i) temporal aspects (tweets from August 2015 to March 2016), (ii) geo-location (tweets within the United States), and (iii) relevant terms derived from our lexicon.

⁶<http://wiki.dbpedia.org/>

(a) Statistics of Cannabis vs. other types of Marijuana

(b) Statistics of the Rest 18% of marijuana types.

Fig. 2: Cannabis and Marijuana mentions statistics obtained from the tweets. Cannabis and Marijuana are the most popular terms within the tweets.

- 2) *Data Processing and Analysis*: This module has two purposes. First, it predicts public opinion (as measured by the sentiment in tweets) at the state level, which requires knowing a user's approximate location and an approach for sentiment prediction. Second, it tracks trends at the state level for the various categories or types of marijuana. To accomplish this, we perform entity extraction to identify the types of marijuana which appear in the tweets collected.
- 3) *Results from our experiments*: This module provides representation for all the data processing and the analysis performed in the previous module to facilitate comparison as well as interpretation.

IV. DATA COLLECTION METHODS

In this section, we present the principles employed in collecting our dataset.

The social media events effect everywhere, users share their opinion even if they don't officially vote. Tweets from all states have shown the users care about the marijuana ballot issues. By having users tweets we could predict which states show more interest to be legalized and even if there is no event in the

state, still people share their ideas. Therefore, for prediction analysis, we collected the data from all US states.

The *Twitris* platform [19] was used to collect and filter Twitter data available through Twitter's streaming API. Twitris filters out non-English language tweets and uses keywords and blacklist words to extract tweets of interest.

Twitris⁷ technology is also commercialized for a startup spin-off. Furthermore, Twitris can be used to perform various analyses on collected tweets (e.g., topic modeling). To collect as many relevant tweets as possible, we were required to develop a comprehensive thesaurus for marijuana. Before elaborating on the details of creating a marijuana thesaurus, we review two further filtering considerations:

- *Data preprocessing*: The duplicates data has removed from the dataset.
- *Temporal filtering*: The polling for marijuana legalization in Ohio was done on November 5, 2015. Thus, we collected relevant tweets from August 5, 2015 (i.e., before marijuana legalization) to November 5, 2015. Furthermore, we collected relevant tweets from November 6, 2015, to March 6, 2016 (i.e., after marijuana legalization). In total, we collected 7.5 million tweets over a period of eight months.
- *Locational filtering*: We collected all the relevant tweets that were published within *the United States*, and to do so, we captured the geo-location tag of tweets and restricted our collection to only those tweets which the Twitter API gave as originating within the US.

Marijuana Thesaurus: CITAR (Center for Interventions, Treatment and Addictions Research) and Kno.e.sis (Ohio Center of Excellence in Knowledge-enabled Computing) created a marijuana thesaurus by collecting 153 keywords that are used interchangeably for marijuana from Twitris-eDrugTrends.

This thesaurus was expanded by including synonyms and the slang terms which are prevalent on social networks. Moreover, Table 1 provide a sample of tweets corresponding to slang terms like *THC*, *blunt*, *spot*, *mj*, etc. which represents marijuana in Twitter.

Entity Extraction Techniques: CITAR and the Kno.e.sis Center jointly developed the *Drug Abuse Ontology (DAO)*⁸. To the best of my knowledge, it is the first ontology on the drug abuse that we have developed for this analysis. New version of DAO is developed to extract the entities semantically and to raise representations of drug mentioned in the tweets.

DAO was initially used for the PREDOSE⁹ project with basic level of drug names which now, developed the techniques to facilitated our prior work on prescription drug abuse epidemiology. The first version of DAO contained 87 classes, the updated classes contained 243 classes (such as *drug*, *dose*, *drug abuse treatment*, and *medical conditions*) and 36 properties (such as *diagnosis*, *causes*, and *interacts*). DAO

⁷<http://twitris.knoesis.org>

⁸<http://wiki.knoesis.org/index.php/DAO>

⁹<http://wiki.knoesis.org/index.php/PREDOSE>

Fig. 3: Hierarchy of Various Marijuana Types from the Drug Abuse Ontology(DAO).

Categories	Slang Terms	Tweets
Marijuana (Cannabis)	hash joint 420 spot spice THC mj	no1. hash made my day, do you like to try it, its better than mj cookies no2. Smoke a joint and get lost in the sound of the ocean waves no3. 420 use to be police code in Call for smoking marijuana in progress. Ppl heard it & started saying let's go 420 lol no4. Check out this amazing smoke spot keep on tracking 420!... no5. daily Spice, smoke and shifters ..@...187 gangsters: Paranoid from Spice!:-)Don't smoke spice! fuckSpice follow. no6. @...People who smoke pot on RNR could not because of urine testing showing THC in samples no7. @...DanyBoy version 8 in the house smoking some mj
Synthetic Marijuana	stoner, weed K2, dro spice ISO	no8. Pol being a stoner doesnt explain the list of anal people like this weed prolly keeps them. no9. I dont smoke that K2 you smoking dro now ...mom just walked in my room and said "dont smoke K2!" no10. daily Spice, smoke and shifters ..@...187 gangsters: Paranoid from Spice!:-)Don't smoke spice! fuckSpice follow ... no11. ISO: 2x spacious, quiet, smoke free section to come join
Edible	mj cookie	no15. @... Mj really likes the birthday cake Oreos. he asks if he can have a happy birthday cookie lmaoo
Marijuana Concentrate	dabs	no16. Night time Rosin dabs and work! https://t.co/LP5RTXWIT8
Marijuana Oil	cbd oil	no17.Ganjavisst @... osalindSt0ne CBD oil works wonders Better balanced weed would probably achieve the same :)
Marijuana Resin	keif keef hashish	no12. @...Buddy gave me a free gram and some keif to top it off with!... no13 noChan @ChandlerNashh I got about a blunt s worth of keef no14. Healthy Vegi Meals @... Mix ice, hashish , and protein powder then put it in the oven.

TABLE I: Samples of relevant tweets about marijuana (Tweet handles were removed for privacy preservation w.r.t. IRB).

has also been enriched by linking to DrugBank, Freebase, DBpedia, and the cognitive-labs¹⁰ knowledge bases.

As part of the full ontology, DAO contains a comprehensive set of slang terms associated with each type of medical marijuana and also provides a useful hierarchy which is represented in Figure 3. In this current study, we use the relevant portion of DAO for entity extraction and obtaining slang terms related to the different types of marijuana.

Filtering on Seed Terms: We gleaned a set of seed terms from the marijuana thesaurus and the slang terms associated with the various types of non-medical marijuana from eDrugTrends campaign that created for this study in Twitris. In conjunction with legalization-related terms, this seed set of marijuana terms was employed for filtering the tweets. Table II provide samples of some of the tweets containing at least one term from our set of seed terms.

Statistics on Collected Data: Please note that we were only collecting the tweets in the English language. After applying our filter, we were left with 4,307,389 relevant tweets out of the 7.5 million which we collected initially. The important statistic is related to the entity identification task on the collected tweets. We identify entities referencing marijuana

and its various types/categories as demonstrated in the DAO ontology (shown in Figure 3). Figure 2 represents statistics on this task. As can be observed in Figure 2(a), 82% of entities reference marijuana and the remaining 18% reference other types of marijuana (shown in Figure 2(b)). Figure 2(b) reveals that among the sub-typed entities concentrates form the highest percentage (73%) and cannabis resin, synthetic cannabis, cannabis oil, and edibles respectively form 12%, 2%, 4% and 9% of the identified entities. Figure 4 illustrates the number of relevant tweets per state for three different categories: (i) Category 1 - states in which no form of marijuana legalization has taken place (shown in Figure 4(a)), (ii) Category 2 - states in which both medical and recreational marijuana have been legalized (shown in Figure 4(b)), and (iii) Category 3 - states in which only medical marijuana has been legalized (shown in Figure 4(c)). For each of these categories, statistics were generated from relevant tweets collected in the *pre-marijuana legalization* period (four months before voting took place) and *post-marijuana legalization* period (after November 5, 2015).

Category 1 represents 21 states and contains 1,548,163 relevant tweets. In this category, 51.93% of the relevant tweets belong to the pre-phase, and 48.06% belong to the post phase. Category 2 represents four states and 445,421 relevant tweets.

¹⁰<https://cognovillabs.com>

In this category, 30.06% of the relevant tweets belong to the pre-phase, and 69.93% belong to the post phase. Category 3 represents 26 states and contains 2,313,805 relevant tweets. In this category, 31.71% of the relevant tweets belong to the pre-phase, and 68.28% belong to the post phase.

(a) Category 1: Neither medical nor recreational legal status.

(b) Category 2: Medical and recreational legal status.

(c) Category 3: Medical legal status.

Fig. 4: Statistics of relevant tweets categorized based on legal status of marijuana per state in the US before and after the election on November 5th 2015.

V. ANALYSIS USING MINING METHODS

Our experimental study is divided into (i) the task of measuring public opinion by determining the majority sentiment associated with the collected tweets at the state level and (ii) the work of showcasing the trends in marijuana consumption (i.e., identified entities) in the managed tweets. Each task is discussed separately in the following subsections.

A. Measuring Public Opinion

The Ohio Marijuana Legalization ballot was on November 3, 2015, where it was defeated (63.65% was "No" and 36.35% was "Yes").

In this study, if the percentage of positive sentiment is higher, it indicates people's positive opinion about marijuana and vice versa. We employed the sentiment analysis algorithm presented in [5] with an external learning module. To select

a well-performing supervised learning module, we had to perform a comparative study on a given training dataset and to prepare our labeled dataset; we annotated 8,450 relevant tweets from our relevant subset of tweets. The annotation task labeled the sentiment (positive, neutral, and negative) implied by a given tweet in our relevant subset. We employed four annotators for annotating all the tweets. The inter-annotator agreement rate was 85% defining fineness in the annotation.

The sentiment algorithm has two lexicons list with negative and positive words. We integrated our sentiment analysis algorithm [5] with four supervised learning algorithms, namely, Logistic Regression (LR) [15], Naive Bayes (NB) [20], ZeroR [8], and a Support Vector Machine (SVM) [8]. These algorithms are known to perform well for sentiment classification task.

SVM is a statistical supervised machine learning technique. The Binary linear SVM classification obtains the calculation of the optimal hyperplane decision boundary, it separates one class from the other, on the basis of a training data-set.

As can be observed, the SVM algorithm outperforms the other supervised algorithms thus, we rely on SVM as the backbone of our supervised approach for measuring sentiment. Table II shows the performance results of employing these four algorithms on our labeled data-set with 10-fold cross-validation.

Algorithm	Precision	Recall	F-Score
Logistic Regression	0.8076	0.8026	0.7964
NB	0.8668	0.7626	0.7992
ZeroR	0.8854	0.8242	0.8484
SVM	0.914	0.889	0.899

TABLE II: Comparison of accuracy for supervised learning algorithms in sentiment analysis.

State	% +ve Sentiment	State	% +ve Sentiment
1. California**	14.48	6. Washington***	5.1
2. Texas*	7.41	7. Ohio*	3.94
3. New York**	6.59	8. Oregon***	3.34
4. Florida*	6.12	9. Michigan**	3.07
5. Colorado***	6.08	10. Illinois**	2.76

TABLE III: The top ten states with the highest percentage of positive sentiment.

a) *Results of Mining Public Opinion.*: We set up our prediction module using the SVM algorithm integrated with the sentiment analysis algorithm mentioned earlier. We ran this prediction module on all of the relevant tweets collected. Table III represents the predicted sentiment per state. The results are represented in the three categories (shown in Figure 5): (i) states in which no legalization has taken place (shown in Figure 5(a)), (ii) states in which medical marijuana has been legalized (shown in Figure 5(b)), and (iii) states in which both medical and recreational marijuana have been legalized (shown in Figure 5(c)). In all the categories the neutral sentiment is dominant (i.e., 65% for the category one, 66% for the category two, and 24% for the category three). The reason behind this dominance is the majority of the tweets about marijuana legalization posted by media and retailers which correlate with public opinion and the other two polarities of sentiment

(a) Category 1: Neither medical nor recreational legal status.

(b) Category 2: Medical legal status

(c) Category 3: Medical and recreational legal status.

Fig. 5: Statistics of relevant tweets categorized based on legal status of marijuana per state in the US before and after the election on November 5th 2015.

(positive and negative) are better indicators of the opinion [5]. The sentiment analysis for categories 1 and 2 represent higher positive sentiment rather than negative sentiment (i.e., 23% positive versus $\approx 11\%$ negative), which can be interpreted as the positive public opinion about marijuana legalization. Conversely, category 3 shows 67% positive sentiment (almost three times more than the other two categories) versus 9% negative sentiment. The significant difference is the legalization of recreational marijuana in this category which can be interpreted as an indicator of higher happiness compared to the other two categories.

Table III shows the top ten states representing the highest percentage of positive sentiment. One asterisk indicates no legalization, two asterisks indicate medical legalization, and three asterisks indicate legalization for both medical and recreational usage. More importantly, note that these statistics are for tweets before March 2016. We hypothesize that the states with the highest positive sentiment indicate that the majority of public opinion is favorable towards legalization.

These observations confirm our hypothesis that positive public opinion measured by social media can be leveraged as an indication of intentions to legalize marijuana for both medical and recreational purposes.

B. Trending Marijuana Types

The second major branch of our analysis is about the trending consumption of distinct types of marijuana in the entire corpus. As we mentioned above, we distinguish six types for marijuana, five of them originating from the medical or recreational categories for marijuana, namely, cannabis, cannabis oil, cannabis resins, edible cannabis, and marijuana concentrates. The sixth type is *synthetic cannabis*, which is used solely in a recreational context.¹¹ We initially represent the consumption distribution of each type of marijuana per state (shown in Figure 6) during the pre-legalization phase (shown in Figure 6(a)) and the post legalization phase (shown in Figure 6(b)). The overall distribution in the pre and post stages can be divided into two trends: (1) an increasing trend which was observed for marijuana concentrates from 30.26% to 69.73% and for cannabis from 46.72% to 53.27%; and for edibles from 44.54% to 55.45%; and cannabis oil from 38.82% to 61.17%; and (2) a decreasing trend observed only in cannabis resin from 51.66% to 48.33%. The increase in the type of cannabis oil can be attributed to the higher availability of it in the market, which adds to its popularity.

Another insight into the trends for the categories of marijuana can be drawn by analysis of the popularity of each type at the state level. In the following, we present the top ten states for each category of marijuana in descending order of popularity. Note that the asterisk represents the legal status for each type of marijuana in this state. One asterisk indicates no legalization, two asterisks indicate only medical legalization, and three asterisks indicate legalization of both medical and recreational usage.

We prepared datasets for each type of marijuana by taking randomly sampled tweets for that particular type from the entire corpus. After preparing the five type-specific datasets (we focused on cannabis types only and we just, we asked our annotator to annotate the given tweets concerning the relatedness to the *consumption of type X of marijuana* using the labels 'yes', 'no' or 'can not be decided.' After that, we individually trained a classifier to predict the consumption of each type. We applied cross-validation on all underlying datasets in an incremental manner (meaning varying the size of the dataset by injecting an increasing number of tweets – starting from 100 tweets and going up to a total of 2,000 tweets).

To find the best classifier, we compared the accuracy of four different algorithms. Table IV shows the details of the accuracy of the results of running the classifiers, namely, (i) Naive Bayes (NB), (ii) Support Vector Machine (SVM), (iii) J48, and (iv) Random Forest (RF) for each type of marijuana. Classifier NB relies on a frequency-based statistics. J48 and RF are dependent on a covariance matrix, and SVM uses similarity measures [9].

This accuracy can be attributed to the strategies that we employed for entity identification task aided by the use of a comprehensive ontology (DAO ontology) and lexicon. Specifically, these strategies enhanced both the recall as well as the precision for recognizing entities (i.e., marijuana types) in social media with informal and noisy language. The small difference in the mean of accuracy for the types such as edible and marijuana concentrate is due to a higher variety of slang terms (which is yet to be captured well in our ontology) used in social media for marijuana concentrate in comparison to that for edible.

VI. RELATED WORK

Marijuana usage and legalization have been the subject of many recent studies. For example, there has been ground-work concerning the initiation of marijuana use among high school seniors after its legalization. [17] report that after legalization there was a 10% increase in the intention to use cannabis among high school seniors, and this attributes to three characteristics: (1) the demographics of the subjects, (2) the substance used (type of drug), and (3) the disapproval of drug use. Traditionally, information concerning marijuana legalization has mainly been collected from news articles and government documents. However, since there are also various popular forms of digital media such as videos, blogs, and microblogs, by omitting these sources from data collection we needlessly limit ourselves from the full spectrum of media platforms which people use to disseminate information and share their opinions. These platforms are an especially useful as sources of information on public health and communication.

Observing the effectiveness of mining social media has motivated us to analyze trends in drug consumption, drug abuse, and perform opinion mining using Twitter. The subject of analyzing social media concerning drugs such as marijuana

¹¹https://en.wikipedia.org/wiki/Synthetic_cannabinoids

Marijuana Types	NB			SVM			J48			RF		
	P	R	F	P	R	F	P	R	F	P	R	F
Marijuana Oil	0.830	0.610	0.710	0.800	0.790	0.790	0.776	0.770	0.772	0.820	0.830	0.826
Marijuana Concentrate	0.802	0.757	0.772	0.274	0.524	0.360	0.715	0.700	0.697	0.681	0.680	0.680
Marijuana Resin	0.436	0.660	0.525	0.699	0.710	0.701	0.770	0.772	0.772	0.827	0.830	0.826
Edibles	0.874	0.875	0.874	0.436	0.660	0.525	0.715	0.700	0.697	0.681	0.680	0.680
Cannabis	0.837	0.866	0.865	0.851	0.851	0.851	0.715	0.697	0.497	0.681	0.680	0.680

TABLE IV: The accuracy results of four different classifiers for predicting usage of each type of marijuana. (P:Precision, R:Recall, F:F-Score)

(a) Distribution of marijuana consumption per state before legalization.

(b) Distribution of marijuana consumption per state after legalization.

Fig. 6: Distribution of marijuana consumption per State.

has been studied in [4]. "Dabbing" is a form of inhaling marijuana and is extremely dangerous because of its psychological and respiratory effects. The study cited performs a detailed examination of tweets containing "dabbing" and its related terms using a keyword-based search. In [21], tweets posted by adolescents before and after legalization in two states were examined for marijuana-related content. Tweets which were in non-English languages were removed from the examination. The experiment showed 65.6% positive sentiment towards marijuana legalization from a filtered sample of 36,969 tweets created from 71,901 tweets. Of the tweets by adolescents, tweets which had mentions of a parent (36%) reported the need for parental support during the use of marijuana. But, [21] uncovered trends and types in marijuana usage using Twitter. After the legalization of marijuana, there has been a significant change in people's perceptions towards marijuana use. Nationwide examination surveys (1991-2014) comprising people in different age groups (8th, 10th, and 12th graders) found support for marijuana legalization among adolescents. The reason is that they perceive marijuana as a non-harmful drug, whereas people who were not pro-marijuana seen it as a harmful. After the legalization of marijuana, there was an increase in the number of marijuana supporters. "Perceived Harmfulness" declined from 84% (1991) to 53.8% (2014), but remained active among 8th graders, resulting in a 33% decline in its usage. The perceived harmfulness among different age groups before and after marijuana legalization motivated [13] to incorporate social media in assessments of public opinion.

Moreover, the post-marijuana legalization phase has seen an increase in positive sentiments towards marijuana use for recreational and medical purposes. Adolescents mainly have contributed to marijuana supporters. The reason for such behavior is not only attributed to demographics but individual preferences which also play a crucial role in defining the likelihood of cannabis promotion [24]. An individual's response towards cannabis determines his/her dosage amount over the years and the impact of a community (peers). In our approach, we consider the state as a community of people either in favor of marijuana's legalization for recreational or medical usage. "Demographics Pro for Twitter" [3] categorized marijuana supporters and noticed a significant amount of people who were African-Americans.

We extended the state-of-the-art approaches by mining public opinion about the recent legalization of marijuana for all the states in the U.S. Furthermore; we studied consumption trends representing the popularity of various types of marijuana. A deficiency observed in a majority of the related state-of-the-art literature are limitations such as lack of considering the entire types of marijuana. For example, the work discussed in [6] only explores Twitter data on marijuana concentrate, and ignores other types of marijuana. Another work demonstrated in [14] is limited to two terms on synthetic marijuana (a synthetic cannabinoid) "K" and "spice". Their search terms are based on the initial version of DAO. The paper [22] studied state policy on marijuana on Twitter. Our work focuses on sentiment analysis concerning legalization of cannabis, cannabis

oil, synthetic cannabis, cannabis resins, edible cannabis and marijuana concentrates six major drug abuse types which were the focus of *dabs* in [7] as *dabs* is a sub-category of marijuana concentrates. Tweets often contain only a short amount of text, but still may provide sentiments regarding a particular target.

VII. CONCLUSION AND FUTURE PLAN

This work mined public opinion on marijuana legalization and consumption trends using a corpus created from Twitter data filtered in both state-wise and temporally. Collecting relevant data with high recall was our primary concern due to challenges posed by the informal language used in Twitter. To address this issue, we employed a lexicon compiled from multiple resources and the Drug Abuse Ontology(DAO).

Our sentiment analysis and consumption trends were performed on a corpus of 306,835 tweets from 4,307,389 relevant tweets (marijuana and legal*) out of 7.5M data, collected over four months preceding November 2015 Ohio Marijuana Legalization ballot and four months after, for all states of the US. We mined public opinion by measuring sentiments attached to the tweets in our subjective corpus. We compared the sentiments measured preceding the election to the feelings measured after. Compelling insights were revealed, such as states with high level of positive sentiment preceding the election were engaged in enhancing their legalization status. In fact, people resided in states that have legalized recreational marijuana express greater positive sentiments about marijuana than the people resided in states that have either only legalized medicinal marijuana or have not legalized marijuana at all. Furthermore, the states that have a high percentage of positive sentiment about marijuana have higher interest to legalize (e.g., by allowing medical marijuana) or broaden its legal usage (e.g., by allowing recreational marijuana in addition to medical marijuana). Moreover, We built individual classifiers with high accuracy, exploiting DAO ontology and lexicon, to determine and analyze the consumption trend for each type of marijuana. These classifiers run with accuracy higher than 80% which is due to the strategies (e.g., using DAO ontology) employed for entity identification task. Thus, by using these classifiers, we can easily monitor the consumption trends in the USA.

In future, we plan to extend our work in five directions: First, improving the ontology to better deal with slang terms appearing in social media. Second, developing a word sense disambiguation methodology for reliable interpretation of marijuana terms such as K2, dabs, and spice. Third, implementing a classifier to differentiate provenance of tweet posts about marijuana, e.g., media, retailers, and advertisers. This differentiation can further improve the accuracy of our analysis. Fourth, applying network analysis techniques to neutral tweets (i.e., tweets without polarity) which is typically published by media, retailers, advertisers to recognize the emerging "topics" or "trends" respecting the subjectivity of marijuana. Fifth, improve DAO by following ontology methodologies and best practices and encourage its re-usability and dissemination.

a) Acknowledgment: We acknowledge support from the National Institute on Drug Abuse (NIDA) Grant

No.5R01DA039454-03: "Trending: Social Media Analysis to Monitor Cannabis and Synthetic Cannabinoid Use". Technology enhancement and scaling of Twitris is funded in part through the National Science Foundation award IIP 1542911: "PFI: AIR-TT: Market-driven Innovations and Scaling up of Twitris - A System for Collective Social Intelligence." Thanks to "Manas Gaur" for the careful review of the paper.

REFERENCES

- [1] Catherine Bartlett and Rebecca Wurtz. Twitter and public health. *Journal of Public Health Management and Practice*, 2015.
- [2] Amelia Burke-Garcia and Gabriel Scally. Trending now: future directions in digital media for the public health sector. *Journal of Public Health*, 2014.
- [3] Patricia A Cavazos-Rehg, Melissa Krauss, Sherri L Fisher, Patricia Salyer, Richard A Grucza, and Laura Jean Bierut. Twitter chatter about marijuana. *Journal of Adolescent Health*, 56(2):139–145, 2015.
- [4] Patricia A Cavazos-Rehg, Shaina J Sowles, Melissa J Krauss, Vivian Agbonavbare, Richard Grucza, and Laura Bierut. A content analysis of tweets about high-potency marijuana. *Drug and alcohol dependence*, 2016.
- [5] Lu Chen, Wenbo Wang, Meenakshi Nagarajan, Shaojun Wang, and Amit P Sheth. Extracting diverse sentiment expressions with target-dependent polarity from twitter. 2012.
- [6] R Daniulaityte, R Carlson, F Golroo, S Wijeratne, E Boyer, S Martins, R Nahhas, and A Sheth. "time for dabs": Analyzing twitter data on butane hash oil use. 2015.
- [7] Raminta Daniulaityte, Lu Chen, Francois R Lamy, Robert G Carlson, Krishnaprasad Thirunarayan, and Amit Sheth. "when 'bad' is 'good'": Identifying personal communication and sentiment in drug-related tweets. *JMIR Public Health and Surveillance*, 2(2), 2016.
- [8] Rebeen Ali Hamad, Saeed M Alqahtani, and Mercedes Torres Torres. Emotion and polarity prediction from twitter. 2017.
- [9] Danesh Irani, Steve Webb, Calton Pu, and Kang Li. Study of trend-stuffing on twitter through text classification. In *Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, 2010.
- [10] Jared Jashinsky, Scott H Burton, Carl L Hanson, Josh West, Christophe Giraud-Carrier, Michael D Barnes, and Trenton Argyle. Tracking suicide risk factors through twitter in the us. *Crisis*, 2014.
- [11] Janet Joy, Alison Mack, et al. *Marijuana as medicine?: the science beyond the controversy*. National Academies Press, 2000.
- [12] David Jurgens, Mohammad Taher Pilehvar, and Roberto Navigli. Cross level semantic similarity: an evaluation framework for universal measures of similarity. *Language Resources and Evaluation*, 50(1):5–33, 2016.
- [13] Katherine M Keyes, Melanie Wall, Magdalena Cerdá, John Schulenberg, Patrick M O'malley, Sandro Galea, Tianshu Feng, and Deborah S Hasin. How does state marijuana policy affect us youth medical marijuana laws, marijuana use and perceived harmfulness: 1991–2014. *Addiction*, 111(12):2187–2195, 2016.
- [14] Francois R Lamy, Raminta Daniulaityte, Ramzi W Nahhas, Monica J Barratt, Alan G Smith, Amit Sheth, Silvia S Martins, Edward W Boyer, and Robert G Carlson. Increased in synthetic cannabinoids-related harms: Results from a longitudinal web-based content analysis. *International Journal of Drug Policy*, 44:121–129, 2017.
- [15] Nasir Naveed, Thomas Gottron, Jérôme Kunegis, and Arifah Che Alhadi. Bad news travel fast: A content-based analysis of interestingness on twitter. In *Proceedings of the 3rd International Web Science Conference*, page 8. ACM, 2011.
- [16] Brendan O'Connor, Ramnath Balasubramanyam, Bryan R Routledge, Noah A Smith, et al. From tweets to polls: Linking text sentiment to public opinion time series. *Icwsn*, 11(122-129):1–2, 2010.
- [17] Joseph J Palamar, Danielle C Ompad, and Eva Petkova. Correlates of intentions to use cannabis among us high school seniors in the case of cannabis legalization. *International Journal of Drug Policy*, 25(3):424–435, 2014.
- [18] Sujan Perera, Pablo Mendes, Amit Sheth, Krishnaprasad Thirunarayan, Adarsh Alex, Christopher Heid, and Greg Mott. Implicit entity recognition in clinical documents. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 228–238, 2015.
- [19] Amit Sheth, Ashutosh Jadhav, Pavan Kapanipathi, Chen Lu, Hemant Purohit, Gary Alan Smith, and Wenbo Wang. Twitris: A system for collective social intelligence. In *Encyclopedia of Social Network Analysis and Mining*, pages 2240–2253. Springer, 2014.
- [20] Songbo Tan, Xueqi Cheng, Yuefen Wang, and Hongbo Xu. Adapting naive bayes to domain adaptation for sentiment analysis. In *European Conference on Information Retrieval*, pages 337–349. Springer, 2009.
- [21] Leah Thompson, Frederick P Rivara, and Jennifer M Whitehill. Prevalence of marijuana-related traffic on twitter, 2012–2013: a content analysis. *Cyberpsychology, Behavior, and Social Networking*, 18(6):311–319, 2015.
- [22] Jason Turner and Mehmed Kantardzic. Geo-social analytics based on spatio-temporal dynamics of marijuana-related tweets. In *Proceedings of the 2017 International Conference on Information System and Data Mining*, pages 28–38. ACM, 2017.
- [23] Stanley J Watson, John A Benson, and Janet E Joy. Marijuana and medicine: assessing the science base: a summary of the 1999 institute of medicine report. *Archives of General Psychiatry*, 57(6):547–552, 2000.
- [24] Jenny Williams, Jan C Van Ours, and Michael Grossman. Why do some people want to legalize cannabis use. Technical report, National Bureau of Economic Research, 2011.