Drugs or Dancing? Using Real-Time Machine Learning to Classify Streamed "Dabbing" Homograph Tweets

Antonio A. Ginart, Sanmay Das, Jenine K. Harris, Roger Wong, Hao Yan, Melissa Krauss, Patricia A. Cavazos-Rehg
Washington University, St. Louis, MO

***Abstract*—Dabbing is a new and popular method of using marijuana that involves inhaling vapors from heating marijuana concentrates. As the emergence of legal, regulated markets continues in the U.S., it is possible that dabbing marijuana concentrates will gain traction. Dabbing may present new hazards to marijuana users including increased risk of fires from igniting extracts with butane and increased incidence of addiction due to higher concentrations of the psychoactive chemical tetrahydrocannabinol (THC) inhaled when dabbing. Twitter can be used to better understand health behaviors by analyzing conversations around marijuana dabbing, however, collecting relevant tweets is complex given that "dabbing" is also a term used to describe a dance done at sporting events and the process of covering a sneeze. We developed a machine learning algorithm to classify tweets and identify relevant marijuana dabbing (mdab) tweets. We found our classifier to be reliable in differentiating mdab from other dabbing tweets. Machine learning based classifiers have potential for helping public health researchers and practitioners to handle the large volumes of complex Twitter data in order to learn from this new information stream. Our technique, used to solve this particular tweet differentiation problem, is easily applicable to any homograph differentiation problem in tweet space.**

***Keywords* - cannabis; dab; marijuana; machine learning; Twitter; homograph**

I. Introduction

Twitter is a microblogging tool with 310 million monthly active users [1] who compose and send more than 500 million tweets per day globally [2] including tweets about health and health behaviors [3, 4]. In the U.S., Twitter is most popular among young adults; 32% of Twitter users are 18 and 29 years old [5]. Young adulthood is also when health behaviors that can influence lifelong health begin to develop. One health behavior that often occurs during young adulthood is marijuana use. Marijuana use is widespread across the U.S. In the 2014 National Survey on Drug Use and Health, 6.8 million young adults between the ages of 18 and 25 years were current users of marijuana [6]. Marijuana use is most prevalent among young adults, with nearly 53% of 18 to 25-year-olds using marijuana in their lifetime and 19.6% using marijuana in the past month [6].

Although individuals have traditionally ingested marijuana through joints and bongs, a new and popular method of using marijuana is "dabbing", which involves inhaling vapors from heating marijuana concentrates [7]. Marijuana concentrates are made by extracting the psychoactive chemical tetrahydrocannabinol (THC) from marijuana plants using a solvent such as butane or carbon dioxide. The product is a sticky substance called wax, shatter, budder, or oil, with very high concentrations of THC, ranging up to 80% [8, 9]. Dabbing marijuana concentrates may present new health hazards including an increased risk of burns and fires from igniting extracts with butane [10, 11] and increased incidence of addiction due to substantially higher concentrations of the [12] ingested when dabbing. Given existing rates of marijuana use and recent legislation that has increased legal accessibility of marijuana, there is a need to better understand the impact of dabbing on the health of young adults [13].

High rates of use of Twitter in the age group most likely to initiate and use marijuana makes Twitter a potentially useful source of data on dabbing for researchers in health and public health. Research on the discussion of marijuana dabbing on Twitter, however, is complex. The term dabbing is also colloquially used to describe a type of dance popular at sporting events. Given the two uses of dabbing and the enormous volume of tweets posted regularly on Twitter, traditional public health research methods used to collect and examine a useful set of dabbing tweets ranges from inefficient to impossible for most researchers. Tools in computer science, particularly machine learning, have the capacity to help public health researchers overcome these challenges [14], as they have in other domains like database construction in the biomedical sciences [15] . In this study, a multi-disciplinary team from medicine, public health, and computer science collaborated to build a classifier to differentiate between tweets pertaining to marijuana dabbing (mdab) and those about other types of dabbing. Specifically, the goal of the project was to develop a classifier to aid in collection and curation of mdab tweets to facilitate learning more about this new health behavior.

II. Data Collection

Using keywords related to dabbing, we collected a large volume of tweets via the Twitter public API. The public API offers two distinct query types, search and streaming. The search query collects recent tweets stored by Twitter based on the query keywords. However, these are not necessarily in any particular order, and tweet popularity may factor into search results. The streaming query collects tweets in real-

time, and thus tweet popularity does not factor into the search results. Streaming queries also use a looser interpretation of the keywords, while search queries more consistently return tweets that strictly contain the keywords. Both query types only allow collection of a small fraction of the total tweets that could match the query keywords.

We collected six sets of tweets, four with search queries and two with streaming queries (see Table 1). No retweets are included in any of the collected data sets (it is simple to filter out retweets with the API). Additionally, the API stamps each tweet with a unique tweet ID. No duplicate tweet IDs were allowed in any of the data sets. A human reviewed the tweets collected, and classified each tweet as relevant or not relevant. Not surprisingly, more specific keywords resulted in a higher percentage of relevant tweets, especially for the search queries (see Table 1, sets 1 and 3) but also resulted in fewer tweets being collected overall and fewer relevant tweets to review than broader search terms.

TABLE I
CHARACTERISTICS OF SIX SETS OF TWEETS COLLECTED VIA TWITTER API USING DABBING-RELATED KEYWORDS

Set	Size	Dates	Query type	Keywords	# of mdab tweets
1	896	2/4 - 2/13	search	(dab OR dabs OR dabbing) AND (weed OR pot OR wax OR smoke OR high OR marijuana OR vape OR oil)	893 (99.67%)
2	6937	2/8 - 2/10	search	dab OR dabs OR dabbing	639 (9.20%)
3	2128	2/8 - 2/17	search	(dab OR dabs OR dabbing) AND (weed OR pot OR wax OR smoke OR high OR marijuana OR vape OR oil)	2127 (99.95%)
4	1266	2/4 - 2/13	search	dab OR dabs OR dabbing	158 (12.48%)
5	2110	2/29 - 3/2	stream	(dab OR dabs OR dabbing) AND (weed OR pot OR wax OR smoke OR high OR marijuana OR vape OR oil)	574 (27.20%)
6	2016	4/8 - 4/12	stream	dab OR dabs OR dabbing	142 (7.04%)

There are several points worth noting about the data. First of all, the characteristics of tweets collected by the search and streaming queries are very different. In particular, the same query yields very different percentages of mdab tweets based on whether it is a streaming query or a search query (compare sets 1 and 3 with set 6). It is possible that the more specific queries also cause us to miss out on particular subsamples of tweets that may be of interest to public health researchers. The streaming queries likely do not suffer from any sample selection bias of this kind. However, the very low percentage of mdab relevant tweets in the streaming query is a problem, since public health researchers could more effectively spend their time manually analyzing tweets if their sample contained a higher proportion of relevant tweets. This motivates the development of a machine learning classifier that can learn to identify relevant tweets with high accuracy.

III. CLASSIFIER IMPLEMENTATION

We used a standard implementation of a linear support vector machine (SVM) from the Python scikit-learn package [16]. We pre-processed each tweet as follows. First, each tweet was standardized by lowercasing the text and stripping it of punctuation and other non-alphanumeric characters. The feature space was then the vector space indexed by all "word tokens" (the remaining lowercase alphanumeric words) that appeared in any tweet. For any tweet, all word tokens that did not appear would have a value of 0, while those that did would be assigned a value using the term-frequency-inverse-document-frequency (TFIDF) method, where the value is determined by the number of occurrences of the word in the tweet divided by the frequency with which that word appears in all tweets. As discussed above, we are interested in performance on the streaming query with keywords "dab OR dabs OR dabbing" as representative of sampling from the "true" domain. Thus we used set 6 as a validation set to determine the parameters of our learning model, including which of the 5 other data sets should be used to train the classifier (see Fig. 1).

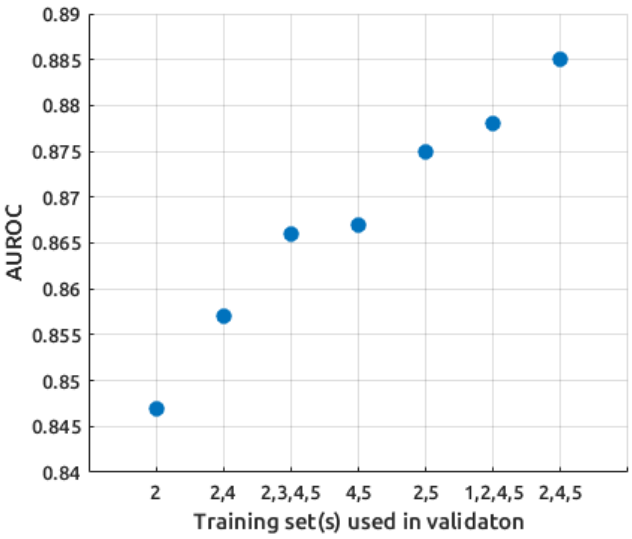

Fig. 1. Validation AUROC for models trained on sets described in Table 1

The next important question is how the SVM should be trained. One design choice is the soft-margin hyperparameter, a regularization parameter that trades off model complexity and training set error. We found that varying the margin parameter C between 0.3 and 1 led to no significant changes in the SVM performance, so we used C=0.5. Another design choice is which combination of training datasets collected above should be used to train the final SVM. This is important because each of the training sets has different characteristics. While the less specific queries are more representative of the distribution we care about, they are problematic in having

many fewer positive examples (relevant to mdabbing). Machine learning methods are known to sometimes have trouble with highly imbalanced datasets.

We evaluate performance using the area under the receiver-operating characteristic curve (AUC). This can be understood as the probability that, if you were to take a randomly chosen truly positive example and a randomly chosen truly negative example, the positive example would be ranked higher by the classifier than the negative example. Fig. 1 shows how well the model worked to differentiate relevant from non-relevant tweets in the validation set, set 6 from Table 1. The highest AUC was for the model trained on sets 2, 4, and 5; after retraining with set 6, we adopted this model as our final classifier.

IV. Results

In order to test the final classifier, we collected test sets A and B, two new sets of 2,500 tweets collected in the exact same manner and during the same timeframe as training set 6. We used these two sets in slightly different manners. The set A was used to evaluate the generalization performance of the classifier using the aforementioned AUC criterion, and to ensure that performance on a held-out test set was comparable to what we would expect. Set B was used to simulate a real use case for public health experts. Each of the test tweets were ranked by the SVM, in decreasing likelihood of mdab relevance. For test set A, all 2,500 tweets were jointly labeled by two members of the project team (they were blind to the classifier predictions and did not label any of the training data) and the AUC was found to be 0.913, indicating strong separation by the SVM of tweets into relevant mdab tweets and tweets that were not mdab related.

The typical use of this kind of classifier would be to set a threshold such that the top k most likely to be relevant tweets (as ranked by the classifier) would be taken as relevant, while the rest would be assumed not relevant and ignored. In practice, a specific regression value could replace an ordinal threshold, and thus the classifier could be applied directly to a real-time stream of tweets. For test set B, the 250 most likely to be relevant tweets (by the classifier's rankings) were labeled by hand by another, third team member (also blind to the classifier's scores). The question is then, what is the precision for a given value of k? So, for example, in both test sets, if the tweet was ranked in the top 25 (1%) most likely to be relevant, then it was actually deemed relevant by human judges 100% of the time (see Fig. 2).

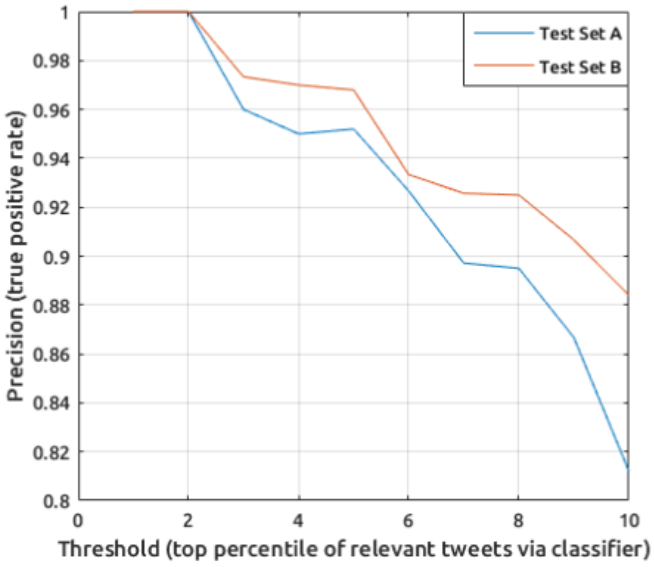

Fig. 2. Precision vs. threshold (k) for test sets

As k increases, there is a tradeoff between getting more relevant tweets to analyze, and having to deal with a larger number of irrelevant tweets in a data set that is supposedly comprised of mdab relevant tweets. Fig. 2 quantifies this tradeoff for the two test sets. Depending on the willingness of the domain expert to deal with irrelevant examples, the threshold could be set to different values.

V. Conclusion and Discussion

We developed and validated a machine learning classifier to differentiate relevant from non-relevant tweets, facilitating research on a new type of marijuana use where the volume and complexity of data collection and coding by hand were prohibitive. The classifier has high precision, especially for the top 10% of tweets identified as most relevant. The techniques applied to this problem should be widely applicable to differentiation of homonyms in tweets.

There are a number of practical benefits to this work. First, consistent with promising early work on collaboration between computer science and public health [14], the classification tool will allow health-focused researchers to collect and examine a large volume of relevant tweets rather than devoting scarce resources to smaller-scale collection and hand-coding of tweets. Second, the building of the tool provided training opportunities for students in public health and computer science to work on an application of new skills outside the classroom and to gain experience in working with a multi-disciplinary team. Finally, the tool developed in this collaborative effort could be implemented by other research teams in health, public health, and other social sciences in order to efficiently collect and curate Twitter data for the purposes of gaining new insights into human behavior, in general, and for the specific study of the use of marijuana in this alternative way. Our tool may be especially useful given the potential for the analysis of large amounts of Twitter data to provide insight that will support public health surveillance of this relatively novel substance use behavior.

References

[1] Twitter. *Twitter Usage*. URL: https://about.twitter.com/company.

[2] R. Krikorian. *Twitter Official Blog: New tweets per second record, and how!* 2013. URL: https://blog.twitter.com/2013/new-tweets-per-second-record-and-how.

[3] J. K. Harris. “Diabetes topics associated with engagement on Twitter”. In: *Preventing Chronic Disease* vol. 12 (2015).

[4] R. G. Tabak L. R. Ruhr J. K. Harris S. Moreland-Russell and R. C. Maier. “Communication about childhood obesity on Twitter”. In: *American Journal of Public Health* vol. 104, no. 7 (2014).

[5] M. Duggan and D. Page. “Mobile messaging and social media 2015”. In: *Pew Research Center* (2015).

[6] Center for Behavioral Health Statistics and Quality. “Behavioral Health Trends in the United States”. In: *Results from the 2014 National Survey on Drug Use and Health*. Substance Abuse and Mental Health Services Administration. Rockville, MD, 2014.

[7] National Institute on Drug Abuse. *What is marijuana?* URL: https : / / www . drugabuse . gov / publications / drugfacts/marijuana#references.

[8] Rachel Chambers. *What is Dabbing and How Do Dabs Work?* 2015. URL: https : / / www . leafly . com / news / cannabis-101/is-dabbing-good-or-bad-or-both.

[9] J.M. Stogner and B.L. Miller. “Assessing the dangers of dabbing: Mere marijuana or harmful new trend?” In: *Pediatrics* vol. 136 (2015).

[10] D. Greenhalgh T. Palmieri G. Jensen R. Bertelotti and P. Maguina. “Honey oil burns: a growing problem”. In: *Journal of Burn Care & Research* vol. 36, no. 2 (2015).

[11] C. J. W. Porter and J. R. Armstrong. “Burns from illegal drug manufacture: Case series and managemen”. In: *Journal of Burn Care & Research* vol. 25, no. 3 (2004).

[12] D. Slade H. Denham S. Foster A. S. Patel S. A. Ross I. A. Khan Z. Mehmedic S. Chandra and M. A. Elsohly. “Potency trends of 9THC and other cannabinoids in confiscated cannabis preparations from 1993 to 2008”. In: *Journal of Forensic Sciences* vol. 55, no. 5 (2010).

[13] M. Loflin and M. Earleywine. “A new method of cannabis ingestion: the dangers of dabs?” In: *Addictive behaviors* vol. 39, no. 10 (2014).

[14] J. A. Epstein. “Collaborations between Public Health and Computer Science: A Path Worth Pursuing”. In: *American Journal of Public Health Research* vol. 1, no. 7 (2013), pp. 166–170.

[15] Sanmay Das; Milton H. Saier Jr., and Charles Elkan. “Finding Transport Proteins in a General Protein Database”. In: *Proceedings of the European Conference on Principles and Practice of Knowledge Discovery in Databases*. Warsaw, Poland, 2007, pp. 54–66.

[16] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* vol. 12 (2011), pp. 2825–2830.