Finding Needles of Interested Tweets in the Haystack of Twitter Network

Conference Paper · August 2016

DOI: 10.1109/ASONAM.2016.7752273

Qiongjie Tian
Arizona State University

Finding Needles of Interested Tweets in the Haystack of Twitter Network

Qiongjie Tian
Computer Science and Engineering
Arizona State University
Email: qiongjie.tian@asu.edu

Jashmi Lagisetty
Computer Science and Engineering
Arizona State University
Email: jlagiset@asu.edu

Baoxin Li
Computer Science and Engineering
Arizona State University
Email: baoxin.li@asu.edu

***Abstract*—Drug use and abuse is a serious societal problem. The fast development and adoption of social media and smart mobile devices in recent years bring about new opportunities for advancing computer-based strategies for understanding and intervention of drug-related behaviors. However, the existing literature still lacks principled ways of building computational models for supporting effective analysis of large-scale, often unstructured social media data. Part of the challenge stems from the difficulty of obtaining so-called ground-truth data that are typically required for training computational models. This paper presents a progressive semi-supervised learning approach to identifying Twitter tweets that are related to personal and recreational use of marijuana. Based on a small, labeled dataset, the proposed approach first learns optimal mapping of raw features from the tweets for classification, using a method of weakly hierarchical lasso. The learned feature model is then used to support unsupervised clustering of Web-scale data. Experiments with realistic data crawled from Twitter are used to validate the proposed approach, demonstrating its effectiveness.**

I. Introduction

Drug use/abuse is among the serious societal problems in the modern age. According to a 2011 report [1], in the United States alone, illicit drug use costs the society more than $193 billion annually and the number is increasing. The impact is also widespread: In 2013, about 24.6 million Americans 12 years old or older were illicit drug users [2]. Accordingly, a lot of research efforts have been devoted to understanding drug-use-related behaviors and the analysis of potential benefits and limitations of various intervention strategies. A key step in such drug-use-related research is the collection of user behavior data.

Most conventional approaches to user data collection are based on recruitment of participants who would provide inputs to a drug-use-related study, e.g., by answering questionnaires carefully designed to gather various types of behavioral and/or demographical data [3][4]. But there are some well-known limitations in such efforts. For example, the sample size is typically small, as it is in general very costly to involve a large population in such studies. More importantly, such questionnaires in general rely on a participant's explicit recall of his/her drug-use behavior, which could be a limiting factor on its own (e.g., issues like incorrect memory or intentional omission of some facts).

The phenomenal growth of social media and smart mobile devices has led to more and more drug-use-related data appearing online. For example, there are many drug-related discussion groups on Facebook, many drug-use-related questions asked and answered on Yahoo!Answers, as well as many drug-related tweets on Twitter.

Such user-generated social media may be collected at a much larger scale (than an explicit user survey) and thus have the potential of offering realistic insights into understanding of substance-use behaviors, their situational factors, and social contexts. A few recent efforts illustrate this nicely. In [5], Christine Lee *et al.* found that the substance-use related behaviors have similar patterns in data from traditional survey-based approaches and those from social media. In [6], Jennifer Whitehill *et al.* studied the relationship between mobile usage of social networking sites (e.g. Facebook and Twitter) and the alcohol use in a large street festival. In [7], Joris Hoof *et al.* conducted one study on analyzing Facebook profiles to show that some Facebook profile elements can be the indicators of real-life behaviors. In [8], Sarah Stoddard *et al.* examined the influence of young people's social networking behaviors on their alcohol and other drug use.

While having demonstrated to some extent the potential of using social media for substance-use research, these existing efforts also revealed the challenges of building computational models for analyzing largely-unstructured social-media. For example, some user attributes that may be readily available from an explicit survey now need complex inference strategies to figure them out. Further, any approach that relies on training from some labeled dataset cannot be easily extended to large-scale analysis. In this paper, we address some of these challenges in the context of illicit marijuana use and its manifestation on Twitter. Specifically, we propose one semi-supervised approach to studying the user behaviors of the illicit marijuana use using noisy, unstructured and large-scale Twitter data. To our knowledge, this is the first work to study marijuana use behaviors using large-scale Twitter data.

II. Related Works

In this section, we briefly review some related work on study of use of marijuana and other substance, including both traditional methods of recruiting participants and more recent approaches using social media data.

IEEE/ACM ASONAM 2016, August 18-21, 2016, San Francisco, CA, USA 978-1-5090-2846-7/16/$31.00 © 2016 IEEE

A. Participant-recruitment Based Research

Johnston *et al.* conducted follow-up surveys on young adults regarding their behaviors related to drug use in [9]. Similar recruitment based approaches were also used to study the effect of marijuana use in adolescent on their depressive symptoms and IQ development [10] [11].

As noted earlier, these population-survey-based efforts are usually very time-consuming merely for the stage of data collection. Another point to note is that the above-mentioned efforts focused more on finding features or trends from the data rather than developing computational approaches for modeling user behaviors.

B. Research Using Social Media

The non-medical use of Adderall (one psychostimulant drug) among college students using Twitter were studied in [12], where the frequencies, percentages and means were analyzed, and the experiments showed that their findings were similar to traditional survey-based methods. To study the smoking behavior on Twitter, Myslin *et al.* collected tweets from Twitter and performed content and sentiment analysis [13]. Cavazos-Rehg *et al.* also performed content analysis of tweets but with a pro-marijuana Twitter handle (@stillblazingtho) plus the demographics of the handle's followers [14]. Volkow *et al.* reported risks of the recreational use of marijuana like the risk of addiction, effect on brain development, relation to mental illness and so on in [15]. Krauss *et al.* studied the hookah smoking behavior on Twitter in [16]. Leah *et al.* reported their research on how posts on Twitter changed after legalizing recreational use of marijuana in two states [17]. Katsuki *et al.* studied the youth non-medical use of prescription medications (NUPM) on Twitter in order to model the frequency of NUPM-related tweets and identified the illegal access to drug abuse via online pharmacies in [18].

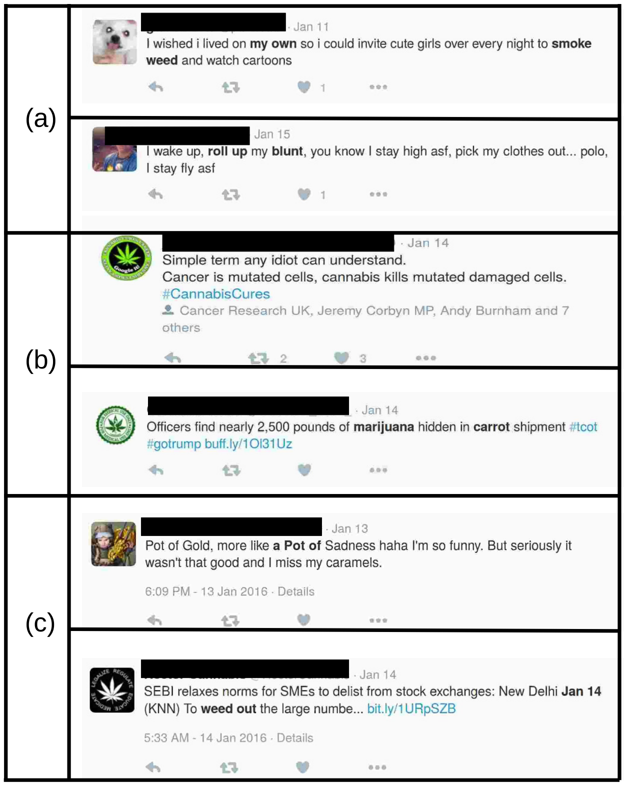
While demonstrating the great potential of using social media for substance-use-related analysis, these existing efforts have yet to be extended to Web-scale data. In particular, we have not seen specific computational models for analyzing Web-scale Twitter data for understanding marijuana-use-related behaviors.

III. Problem Definition

To study the behavior of marijuana users on Twitter, a fundamental problem is to identify tweets that are related to some underlying users who use marijuana. This problem is more subtle than it appears. For example, one cannot simply rely on using the keyword "marijuana" to search the tweets for solving the problem. There are several complicating factors. First, many "street names" are used to describe marijuana. Second, there may be many tweets that involve medical or research-oriented references to marijuana but they are not at all useful for a study on illicit marijuana use. Considering these factors, we propose to classify a tweet into one of following three categories:

- Class One: Tweets in this class are related to personal recreational use of marijuana. They are posted by individual users instead of some official accounts (for example, those for newspaper, companies, or medical institutes).
- Class Two: In this class, all tweets are related to marijuana but not in the sense of recreational use. For instance, they may discuss the medical or prescription use of marijuana, or report some news involving marijuana.
- Class Three: This is for those tweets having no identifiable relationship with marijuana use.

Figure 1 illustrates several real examples for each of the three classes defined above.



[redacted] · Jan 11
I wished i lived on **my own** so i could invite cute girls over every night to **smoke weed** and watch cartoons
1

[redacted] Jan 15
I wake up, **roll up** my **blunt**, you know I stay high asf, pick my clothes out... polo, I stay fly asf
1

[redacted] · Jan 14
Simple term any idiot can understand.
Cancer is mutated cells, cannabis kills mutated damaged cells.
#CannabisCures
Cancer Research UK, Jeremy Corbyn MP, Andy Burnham and 7 others
2 3

[redacted] · Jan 14
Officers find nearly 2,500 pounds of **marijuana** hidden in **carrot** shipment #tcot #gotrump buff.ly/1Ol31Uz

[redacted] · Jan 13
Pot of Gold, more like **a Pot of** Sadness haha I'm so funny. But seriously it wasn't that good and I miss my caramels.
6:09 PM - 13 Jan 2016 · Details

[redacted] · Jan 14
SEBI relaxes norms for SMEs to delist from stock exchanges: New Delhi **Jan 14** (KNN) To **weed out** the large numbe... bit.ly/1URpSZB
5:33 AM - 14 Jan 2016 · Details

Fig. 1: Demos to show three classes: (a) is for Class One, (b) is for Class Two and (c) is for Class Three.

Various text-based features may be extracted for the task of classifying the tweets. Also, as evident from the related work, it is important to consider social interactions among the underlying users. Furthermore, all these features are not mutually independent, and their intricate correlation may provide additional evidence for improved classification. Considering these, and with the goal of classifying large-scale tweets in mind, we now discuss our overall approach, which is illustrated in Figure 2. In the approach, we first extract a set of basic features from each tweet. Then, utilizing a small labeled training set, we learn a good feature mapping that takes into consideration both some basic features *and* their interactions, based on weakly-hierarchical lasso. The learned feature mapping model is used to process the large-scale data.

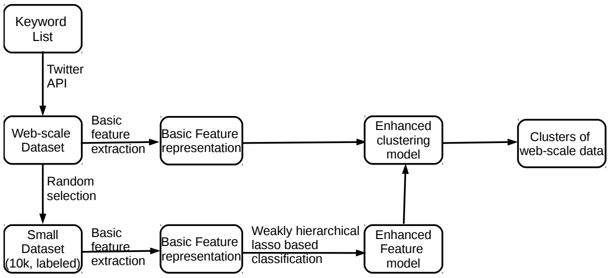

Keyword List
Twitter API
Web-scale Dataset
Basic feature extraction
Basic Feature representation
Enhanced clustering model
Clusters of web-scale data
Random selection
Small Dataset (10k, labeled)
Basic feature extraction
Basic Feature representation
Weakly hierarchical lasso based classification
Enhanced Feature model

Fig. 2: It shows the entire framework of our methodology.

In the following, we first present the basic set of features designed for our task. These features are extracted from either the content of the underlying tweets or the social interactions among the corresponding users, as elaborated below.

A. *Content-based Features*

- The length of the tweet: For each tweet, its length can be one useful feature. For example, the tweets from ordinary users may be generally shorter than those from official accounts.
- Favorite Count & Retweeted Count: It shows how many people think the tweet is favorite and the number people who retweet this post. This is in general useful for measuring how influential the tweet is.
- The number of Hash-Tags: This calculates how many trends one tweet mentions. Our original dataset were obtained by crawling using selected street names of marijuana. The tweets with more trends are likely to be classified as Class Three or Two, instead of Class One.
- TF-IDF on Unigram: Unigram is one common feature used to capture characteristics of one tweet. We build TF-IDF for unigrams of each tweet and use it as one feature.

B. *User-based Features*

- Number of followings and followers: Each user on Twitter can follow others or be followed. However for some official accounts or famous people, they are likely to have a smaller number of followings but a large number of followers. These users are unlikely to post tweets related to personal and recreational use of marijuana.
- Number of Tweets: This records how many tweets one user has already posted, capturing the level of Twitter activity of the user.

IV. Learning Feature Mapping From A Small Dataset

Considering the computational efficiency needed for processing Web-scale data, we may employ a linear classifier as the baseline for doing the classification, as given by Eqn.1.

$$y_i = f(x_i w) \tag{1}$$

where the i^{th} data point is $x_i \in \mathbb{R}^{1\times d}, i \in \{1, \cdots, N\}$ which is normalized, and its label is $y_i \in \{1, 2, 3\}$ and the coefficient to learn is $w \in \mathbb{R}^{d\times 1}$. In this paper, the discriminant function is chosen to be one-vs-one linear SVM. The implementation details are provided as follows. We first train one linear regression model by optimizing Eqn.2.

$$\min_{w} \quad \|Xw - y\|_2^2 + \frac{1}{2}\|w\|_2^2 \tag{2}$$

where $X \in \mathbb{R}^{N\times d}$ and $y \in \mathbb{R}^{N\times 1}$. Then we apply one-vs-one linear SVM to $s = Xw \in \mathbb{R}^{N\times 1}$ to find the label for each tweet.

$$\min_{v} \quad \|v\|_2^2 + C\sum_{i=1}^{N}\xi_i \tag{3}$$

$$\text{s.t.} \quad y_i(s_i * v + b) \geq 1 - \xi_i \quad \forall i \tag{4}$$

where ξ is non-negative.

However, in practice, the linear model is inadequate for capturing the high degree of non-linearity that typically exists in our problem, which has been shown in our experiments. To allow some level of nonlinearity while maintaining computational efficiency, we introduce to the problem 2^{nd}-order interaction terms with a weakly hierarchical structure, as described in [19][20]. The resultant model is given in Eqn.5.

$$\begin{aligned} y &= f(z) \\ z &= xw + \frac{1}{2}\sum_{i}^{d}\sum_{j}^{d} x_i x_j Q_{i,j} \end{aligned} \tag{5}$$

where z is called the z-term of x (for simplicity) and the discriminant function $f(\cdot)$ is given in Eqn.3 (one-vs-one linear SVM in this paper) and x_i is the i^{th} dimension of the data point x and $Q_{i,j} \in R$ is the coefficient for the interaction between i^{th} and j^{th} dimensions of the feature space.

To solve the classification problem under this new model, we formulate the following optimization problem in Eqn.6.

$$\begin{aligned} \min_{w,v,Q} \quad & \frac{1}{2}\sum_{i}(f(z_i, v) - y_i)^2 + \lambda_1\|w\|_1 + \frac{\lambda_3}{2}\|Q\|_1 \\ \text{s.t.} \quad & \|Q_{\cdot,j}\|_1 \leq |w_j| \quad \text{for} \quad j = \{1, \cdots, d\} \end{aligned} \tag{6}$$

where z_i is the z-term of x_i as defined in Eqn.5, $\|Q\|_1 = \sum_{i,j}|Q_{i,j}|$ and v is the model parameter of the discriminant function (the one-vs-one linear SVM).

A. *Solving the Optimization Problem*

Solving Eqn.6) directly is difficult. Hence we simplify this optimization problem by a two-step process: We first learn parameters w and Q and then learn the model parameter v of the discriminant function.

For parameters w and Q, we model them as one regression model as Eqn.7 when we do not consider the discriminant function.

$$\begin{aligned} \min_{w,Q} \quad & \frac{1}{2}\sum_{i}(z_i - y_i)^2 + \lambda_1\|w\|_1 + \frac{\lambda_3}{2}\|Q\|_1 \\ \text{s.t.} \quad & \|Q_{\cdot,j}\|_1 \leq |w_j| \quad \text{for} \quad j = \{1, \cdots, d\} \end{aligned} \tag{7}$$

where z_i is the z-term of x_i as defined in Eqn.5. Then after w and Q are obtained, we learn v of the discriminant function by optimizing Eqn.3.

Converting Eqn.6 into Eqn.7 and Eqn.3 allows us to solve the original optimization problem. By solving Eqn.7 and Eqn.3, we can obtain the model parameters v, w and Q which satisfy the original problem (Eqn.6) as well. However, since we add more constraints on these parameters in the process of simplification, the obtained v, w and Q are only the local optima of Eqn.6.

While the details for solving Eqn.7 can be found in [19], a brief description is given below. From Eqn.7, we can see that this optimization problem is non-convex because of the existence of constraints, and as a result, we cannot solve it using convex optimization approaches. Thus in [19], one convex relaxation by setting $w = w^+ - w^-$ is given, where w^+ and w^- are nonnegative. The convex relaxation version is given as Eqn.8.

$$\min_{w^+,w^-,Q} \quad \frac{1}{2}\sum_i(\hat{z}_i - y_i)^2 + \lambda_1(w^+ + w^-) + \frac{\lambda_3}{2}\|Q\|_1 \quad (8)$$

$$\begin{aligned} \text{s.t.} \quad & \|Q_{\cdot,j}\|_1 \le w_j^+ + w_j^- \quad \text{for} \quad j = \{1,\cdots,d\} \\ & w_j^+, w_j^- \ge 0 \quad \text{for} \quad j = \{1,\cdots,d\} \end{aligned} \quad (9)$$

where $\hat{z}_i = x_i \cdot (w^+ - w^-) + \frac{1}{2}\sum_j^d \sum_k^d x_{i,j}x_{i,k}Q_{i,j}$. A lot of convex optimization approaches can be used to solve Eqn.8, such as FISTA [21].

After we obtain the parameters w and Q, the original problem will become equivalent to the support vector machine which can be solved using sequential minimal optimization.

V. Clustering with The Learned Feature Mapping

A supervised approach cannot be directly applied to Web-scale datasets as manually-labeled data are in general in a much smaller scale. A semi-supervised approach would rely on unsupervised clustering to first identify the structures of the data and then employ a small amount of labeled data to annotate the structures. For example, using K-means clustering, we can group a dataset into different clusters. For data points in each cluster, if we assume that they have the same labels, we can randomly select a small number of data points for labeling and then use the labels to annotate the clusters. Assuming k groups in a dataset, a basic K-means algorithm is equivalent to solving the following problem (Eqn.10):

$$\min_{\pi_j, j\in\{1,\cdots,k\}} \sum_{j=1}^{k}\sum_{v\in\pi_j}\|x_v - c_j\|_2^2 \quad (10)$$

where c_j is the j^{th} centroid and π_j is the j^{th} cluster.

As we have presumably found a feature mapping scheme in the previous section by maximizing classification accuracy for the labelled data, it is natural to use the learned feature mapping for the clustering stage. Denote the dataset as $\{x_i, i \in \{1,\cdots,N\}\}$. Consider the influence of the 2-order feature interaction, the dataset representation is converted as $\{\tilde{x}_i, i \in \{1,\cdots,N\}\}$ where $\tilde{x}_i$ is given by Eqn.11.

$$\tilde{x}_i = (x_i, vec(R_i)) \quad (11)$$

where the element at (j,k) in the matrix R_i is the product of the j^{th} and k^{th} dimension which is $x_{i,j}x_{i,k}$. It is easy to see $\tilde{x}_i \in \mathbb{R}^{1\times(d+d^2)}$. For the new representation, the interaction of the feature dimension is captured by parameters w and Q which are learned from the small labeled dataset (see Section IV). By treating the learned parameters as a kernel, we can have the new clustering as Eqn.12.

$$\min_{\pi_j, j\in\{1,\cdots,k\}} \sum_{j=1}^{k}\sum_{v\in\pi_j}(\tilde{x}_v - c_j)M(\tilde{x}_v - c_j)^T \quad (12)$$

where the learned metric matrix $M = diag((w; vec(Q))) \in \mathbb{R}^{(d+d^2)\times(d+d^2)}$.

VI. Experiments

In this section, we evaluate the performance of our approach with comparison with several typical existing methods.

A. Dataset Construction

For constructing a small labelled dataset, instead of crawling random tweets online, we first use a list of keywords as one filter to remove unrelated tweets. These keywords are defined based on several Web sources and some government documents[1].

The final keyword list was determined to be: *marijuana*, *weed*, *blunt*, *cannabis*, *pot*, *reefer*, *buds*, *420*, *mary jane*, *blaze*. With the final list, the Twitter API [2] is utilized to crawl data. The time period we crawled is from January 09 to January 15 in 2016 and all tweets are in English. We crawled a total of 1,166,441 tweets. Among these we randomly labeled 10,000 with comparable proportion for each class (see Table 1 for exact composition in terms of class labels). This small labelled dataset was annotated by two people reading the tweets to decide their labels.

B. Learning the Feature Mapping

In this part, to compare with commonly used classifiers like linear classifier (Eqn.1) and linear SVM, we split the 10,000 tweets randomly into two parts: training set of 8,000 tweets and testing set with 2,000 tweets. Since in the our approach, we need to compute the feature interaction terms which is defined as the z-term in Eqn.5, we have to reduce the dimension of the original feature vectors. In this experiment, we use LDA [22] to do dimension reduction of TF-IDF of Unigram in the feature sets for our approach. For random guess, we randomly assign one label to every data point and then compute the accuracy based on Eqn.13.

$$e = \frac{\sum_{i=1}^{N_t} I(y_i == \hat{y}_i)}{N_t} \quad (13)$$

[1] In this paper, we use this forum (www.rehabs.com) and this official document(https://vva.org/wp-content/uploads/2014/12/street-terms.pdf).

[2] https://dev.twitter.com/rest/public

where y_i is the ground-truth label of the tweet x_i and $\hat{y}_i$ is the predicted label and

$$I(x) = \begin{cases} 1 & \text{if } x \text{ is true} \\ 0 & \text{otherwise} \end{cases} \qquad (14)$$

The experiment results are shown in Table II and The confusion matrix of our approach is shown in Table I.

RG	*LC*	*SVM*	*Ours*
0.326	0.462	0.677	0.976

TABLE II: The table shows the performance of each baseline and our method. RG: random guess; LC: linear classifier; SVM: linear SVM.

From Table. II, we can easily see that our algorithm stands out. Compared with the modified linear classifier (Eqn. 1 and Eqn. 3) with our algorithm, the difference is that we consider the interaction terms (the z-term) defined in Eqn.5. Thus these results also show that it is necessary to consider feature selection scheme using weakly hierarchical lasso. Furthermore, our approach performs better than linear SVM. This is also easy to understand because of the nonlinearity introduced in our formulation (Eqn. 5). Nonlinearity comes from the z-term.

To further show the performance of each algorithm, the confusion matrices are shown in Table. I. It shows that our approach performs best in all of the three classes. From Table I(a), we can see that *LC* cannot distinguish Class 1 and Class 2. For example, for Class 2, almost the same number of tweets are classified into Class 1 and Class 2. The baseline with *SVM* performs better than *LC*, but the error is still significant.

Our approach effectively solves the problem of how to fuse features and provides the optimal feature selection/combination scheme. It is possible to analyze which features (or their interactions) are most influential. Table III shows the top three main factors which affect the classification performance and their corresponding coefficients. From this table, it can be seen that the number of retweets and also the TF-IDF of Unigram play important roles in distinguish these three classes. We can also see that the content of the tweets is most important for classification. Based on the results of Table III, the top interactions are from the two TF-IDF feature dimensions. This is also demonstrated by the experiment results (see Table IV).

retweet_num	TF-IDF1	TF-IDF2
4.41e-05	4.53e-01	3.62e-01

TABLE III: Illustration of top-3 main factors. The second one and the third one are from TF-IDF of Unigram.

TF-IDF1 * TF-IDF1	TF-ID2 * TF-ID2	TF-ID1 * TF-IDF2
-2.258e-1	1.844e-1	7.884e-2

TABLE IV: This table shows top-3 interaction factors and their corresponding coefficients.

C. Clustering Structure on the Web-Scale Data

In this part, we apply the learned feature mapping scheme to the large dataset, which contains not only the labeled data points but also unlabeled ones. To show the clustering structure of the partially labeled dataset, we perform two experiments: one using one baseline which is KMeans and the other one is our method based on Eqn. 12. For a good clustering outcome, we assume in each cluster, a majority of data points belong to the same class. To evaluate the performance of the results, we present two metrics (Eqn. 15) to show whether any class is dominant in a given cluster. In each cluster, there may be three classes with sizes n_0, n_1 and n_2 (in non-increasing order) respectively. If one class does not exist, it means its size is zero.

$$m_1 = \frac{n_2}{n_0} \qquad m_2 = \frac{n_1}{n_0} \qquad (15)$$

In our experiment, the large dataset is partially labeled and thus when we compute m_1 and m_2, we only consider the labeled data in each cluster. Then the average is computed for the entire dataset. These two metrics are presented to measure what is the difference between the dominant class and the others. If the values of these metrics are small, then they shows that compared with the size of the dominant class, the others are small.

In our experiment, we choose the number of clusters to be $k \in \{10, 100, 200, 300, 400, 500, 1000\}$. In this way, we can learn the effect of the number of clusters on the clustering performance. The experiment results are shown in Table V.

k	$\bar{m}_1$	$\bar{m}_2$
10	0.411	0.647
100	0.320	0.594
200	0.333	0.594
300	0.318	0.602
400	0.291	0.542
500	0.282	0.542
1000	0.239	0.495

(a) The baseline

k	$\bar{m}_1$	$\bar{m}_2$
10	0.381	0.555
100	0.280	0.487
200	0.240	0.436
300	0.263	0.485
400	0.243	0.423
500	0.228	0.427
1000	0.116	0.320

(b) our approach

TABLE V: Experiment results on studying the clustering structure of partially labeled dataset. (a) for the baseline and (b) for ours. They show the size of the other class compared with the dominant one.

From Table V, we can see that our clustering approach by employing the learned feature mapping scheme performs better than the baseline. As the number of clusters goes up, $\bar{m}_1$ and $\bar{m}_2$ of KMeans and our approach become small, which means that the percentage of the dominant class becomes large. Compared with the baseline, the percentage of the dominant class is much larger since the corresponding metrics' values are smaller. The average percentage of the dominant class is shown in Fig.3.

VII. Conclusion and Future Work

We presented one semi-supervised approach to analysis of Twitter data related to marijuana use, using web-scale data,

	C1	C2	C3
C1	0.4616	0.3108	0.2276
C2	0.3820	0.3920	0.2260
C3	0.1751	0.3146	0.5103

(a) *LC*: modified linear classifier

	C1	C2	C3
C1	0.6060	0.1508	0.2432
C2	0.1820	0.6120	0.2060
C3	0.1259	0.0780	0.7962

(b) *SVM*: linear SVM

	C1	C2	C3
C1	0.9831	0.0130	0.0039
C2	0.0020	0.9920	0.0060
C3	0.0014	0.0410	0.9576

(c) *Ours*: our approach

TABLE I: Three confusion matrices for three algorithms: *LC*(a), *SVM*(b), *Ours*(c).

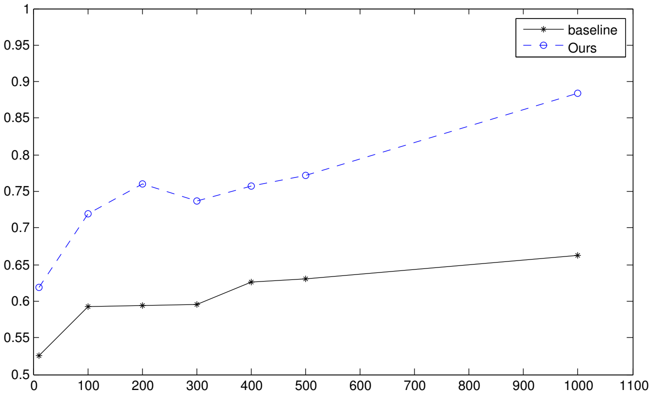


Fig. 3: It shows the average percentages of the dominant class plotted based on the experiment result at each $k \in \{10, 100, 200, 300, 400, 500, 1000\}$

which includes: learning the optimal feature mapping scheme and grouping the entire data using an improved clustering algorithm. The experimental results demonstrated the effectiveness and efficiency of our approach.

There are still some limitations we need to work on. For example, when we learn the feature mapping scheme, we relax the problem to be one easier one, and thus the learned parameters are only locally optimal. Another problem is how to incorporate features reflecting temporal patterns of user behaviors.

Acknowledgment

This work was supported in part by a grant (#1135616) from National Science Foundation. Any opinions expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

[1] U. D. o. J. N. D. I. Center, "The economic impact of illicit drug use on american society," *Product No. 2011-Q0317-002*, 2011.

[2] S. Abuse and M. H. S. Administration, "Results from the 2013 national survey on drug use and health: Summary of national findings," *NSDUH Series H-48*, vol. 14, no. 4863, 2014.

[3] K. J. Quintelier, K. Ishii, J. Weeden, R. Kurzban, and J. Braeckman, "Individual differences in reproductive strategy are related to views about recreational drug use in belgium, the netherlands, and japan," *Human Nature*, vol. 24, no. 2, pp. 196–217, 2013.

[4] J. C. A. Lacson, J. D. Carroll, E. Tuazon, E. J. Castelao, L. Bernstein, and V. K. Cortessis, "Population-based case-control study of recreational drug use and testis cancer risk confirms an association between marijuana use and nonseminoma risk," *Cancer*, vol. 118, no. 21, pp. 5374–5383, 2012.

[5] C. Lee, "Recruitment through social networking sites: Are substance use patterns comparable to traditional recruitment methods?" in *Medicine 2.0 Conference*. JMIR Publications Inc., Toronto, Canada, 2014.

[6] J. M. Whitehill, M. A. Pumper, and M. A. Moreno, "Emerging adults use of alcohol and social networking sites during a large street festival: A real-time interview study," *Substance abuse treatment, prevention, and policy*, vol. 10, no. 1, p. 1, 2015.

[7] J. J. van Hoof, J. Bekkers, and M. van Vuuren, "Son, youre smoking on facebook! college students disclosures on social networking sites as indicators of real-life risk behaviors," *Computers in human behavior*, vol. 34, pp. 249–257, 2014.

[8] S. A. Stoddard, J. A. Bauermeister, D. Gordon-Messer, M. Johns, and M. A. Zimmerman, "Permissive norms and young adults alcohol and marijuana use: The role of online communities," *Journal of Studies on Alcohol and Drugs*, vol. 73, no. 6, pp. 968–975, 2012.

[9] L. D. Johnston, *Monitoring the Future: National Survey Results on Drug Use, 1975-2008: Volume II: College Students and Adults Ages 19-50*. DIANe Publishing, 2010.

[10] R. M. Schuster, R. Mermelstein, and L. Wakschlag, "Gender-specific relationships between depressive symptoms, marijuana use, parental communication and risky sexual behavior in adolescence," *Journal of youth and adolescence*, vol. 42, no. 8, pp. 1194–1209, 2013.

[11] N. J. Jackson, J. D. Isen, R. Khoddam, D. Irons, C. Tuvblad, W. G. Iacono, M. McGue, A. Raine, and L. A. Baker, "Impact of adolescent marijuana use on intelligence: Results from two longitudinal twin studies," *Proceedings of the National Academy of Sciences*, p. 201516648, 2016.

[12] C. L. Hanson, S. H. Burton, C. Giraud-Carrier, J. H. West, M. D. Barnes, and B. Hansen, "Tweaking and tweeting: exploring twitter for nonmedical use of a psychostimulant drug (adderall) among college students," *Journal of medical Internet research*, vol. 15, no. 4, 2013.

[13] M. Myslín, S.-H. Zhu, W. Chapman, and M. Conway, "Using twitter to examine smoking behavior and perceptions of emerging tobacco products," *Journal of medical Internet research*, vol. 15, no. 8, 2013.

[14] P. Cavazos-Rehg, M. Krauss, R. Grucza, and L. Bierut, "Characterizing the followers and tweets of a marijuana-focused twitter handle," *Journal of medical Internet research*, vol. 16, no. 6, 2014.

[15] N. D. Volkow, R. D. Baler, W. M. Compton, and S. R. Weiss, "Adverse health effects of marijuana use," *New England Journal of Medicine*, vol. 370, no. 23, pp. 2219–2227, 2014.

[16] M. J. Krauss, S. J. Sowles, M. Moreno, K. Zewdie, R. A. Grucza, L. J. Bierut, and P. A. Cavazos-Rehg, "Peer reviewed: Hookah-related twitter chatter: A content analysis," *Preventing chronic disease*, vol. 12, 2015.

[17] L. Thompson, F. P. Rivara, and J. M. Whitehill, "Prevalence of marijuana-related traffic on twitter, 2012–2013: a content analysis," *Cyberpsychology, Behavior, and Social Networking*, vol. 18, no. 6, pp. 311–319, 2015.

[18] T. Katsuki, T. K. Mackey, and R. Cuomo, "Establishing a link between prescription drug abuse and illicit online pharmacies: Analysis of twitter data," *Journal of medical Internet research*, vol. 17, no. 12, 2015.

[19] J. Bien, J. Taylor, and R. Tibshirani, "A lasso for hierarchical interactions," *Annals of statistics*, vol. 41, no. 3, 2013.

[20] Y. Liu, J. Wang, and J. Ye, "An efficient algorithm for weak hierarchical lasso," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 283–292.

[21] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[22] Q. Gu, Z. Li, and J. Han, "Linear discriminant dimensionality reduction," in *Machine Learning and Knowledge Discovery in Databases*. Springer, 2011, pp. 549–564.