Identifying Individuals Amenable to Drug Recovery Interventions through Computational Analysis of Addiction Content in Social Media

Ryan Eshleman[1], Deeptanshu Jha[1], and Rahul Singh[1,2]
[1]Department of Computer Science, San Francisco State University, [2]Center for Discovery and Innovation in Parasitic Diseases, University of California, San Diego

***Abstract*—Drug abuse and addiction is a growing epidemic at the forefront of public health. Within this remit, the illicit use of opioid analgesics alone has emerged as one of the fastest growing forms of drug abuse in the U.S. and the death rate from this epidemic are drawing comparison to the US AIDS epidemic. Traditional methods of epidemiology based on explicit reporting of indicator-based data from patient records or data collected through surveys is often found wanting in modeling and designing effective interventions for addiction. In addition to the non-real time nature of the aforementioned methods, this is also due to a number of reasons including the continual penetration of novel biological/chemical entities into the abuse-cycle, the complex etiology of addiction which includes among others social factors, and the multistage nature of the addiction process.**

The recent advent of social media presents an intriguing information resource that is free from some of the above deficiencies and may be leveraged to model the addiction process and offer perspectives that are unavailable through traditional methods of epidemiology. In this paper, we use addiction related social media content to design a computational epidemiological approach for predicting a user's propensity for seeking drug recovery interventions. Solving this problem is crucial for designing effective interventions, identifying cohorts who would be most amenable to recovery, and planning resource allocations. Our method characterizes the evolving language of drug use, identifies the interactions that influence a drug user's actions, and using machine learning techniques predicts the extent to which a user is likely to participate in addiction recovery communities. Experimental assessments on real-world data from the social media platforms Reddit and Twitter indicate the proposed method can identify users who are amenable to addiction recovery intervention with high precision, recall, and F1 values.

Keywords—Computational epidemiology, drug addiction, opioid addiction, addiction analysis, social media

I. INTRODUCTION

In the United States, substance abuse is a major public health concern characterized by rising addiction rates and overdose related deaths [3] with approximately 142 deaths occurring per day [19]. These facts have been met with calls to action by the US government including the White house and Centers for Disease Control [4, 19].

Traditional drug abuse epidemiology has relied on two sources of data: post-factum medical reporting, which occurs once a patient receives medical help, and voluntary surveys. There are significant failings in these methods: first, many patients do not seek timely medical assistance, second, these methods fail to capture the addiction characteristics in real time and finally, they tend not to capture the contextual specificity of affected individuals - which plays a crucial role both in initiation and sustenance of addiction. In this light, social media has received attention as a source of data to fill in these knowledge gaps due to the engaged user base and candid conversation content encouraged by the semi-anonymity of the platforms [5, 6].

Reddit is a forum based social network that boasts 234 million users and has 853,824 unique forums with over 800 million posts in 2015. Users spontaneously form content-centered communities through participation in forums known as *subreddits*. Adherence to the subreddit topic is typically enforced by moderators. Among the diverse subreddits, many focus on recreational drug use and recovery from drug addiction. Our goal is to understand these patterns in a manner that helps guide interventions that not only arrest the spread of addiction but also help the afflicted community. *The specific challenge we seek to address in this paper is the design of predictive models that can identify users who are amenable to recovery assistance.* Given the social media signals we use, we formulate this problem as that of *identifying users who are likely to engage in addiction recovery communities (i.e. are amenable to recovery interventions) based on their activities in drug use communities.*

Solving this problem requires addressing three social media-specific data analysis problems. *First*, algorithms have to accommodate the shifting drug lexicon and adequately represent its content. *Second*, the information organization within a specific social media forum may itself evolve as users create and participate in new subreddits. Finally, a modeling strategy has to be developed which can take the aforementioned inputs and predictively identify users amenable to addiction recovery.

Our approach consists of the following stages: (1) dynamic construction of an addiction lexicon by using term embedding. The embedding space is subsequently mined through a seeded iterative set expansion technique to obtain a representative lexicon. (2) Identification of drug related user communities using this lexicon: subreddits are characterized by the lexicon terms found within and a subreddit network is clustered to identify drug related subreddits. (3) Characterization of individual users based on patterns from drug use subreddits to

978-1-5090-3050-7/17/$31.00 ©2017 IEEE

define a user activity space. A supervised classification formulation is used to predict whether a user will post in an addiction recovery forum. The key contributions of our work are:

1. *Creation of individual-level models to precisely identify substance-users amenable to recovery interventions*. Such models can, among others, ensure that interventions are likely to succeed and that resources are allocated in manners that are most effective.

2. *A dynamic lexicon generation method called Density Based Lexicon Expansion with Seeding.* This method combines distributed vector-space word representations with seeded density-based clustering to represent continually evolving lexicons.

3. *Content-based discovery of characteristic patterns in the drug-use communities.* Among others, we have found that a significant percentage (25.8%) of users in the "opiates" community interact with the "recovery" communities. Hitherto unknown relationships, such as this, can facilitate understanding of community dynamics by characterizing key tendencies and help design interventions.

II. Prior Work

In the context of substance abuse, social media analysis has been pursued in a number of investigations. In the majority, these works have focused on the discovery and characterization of substance use information found in social media. Among others, an investigation into the quantity and quality of information about the drug MDMA/Ecstasy available on the internet was conducted in [9]. The scope of this study was expanded in [5], where a systematic approach was used to scan the web and social media outlets for emerging drug use trends. This method was particularly successful in identifying several new trends, such as the emergence of the drugs *mephedrone* and *spice*. The PREDOSE system [10], was developed to mine social forums for mentions of prescription drug use. In it, the focus on social forums constituted a step toward characterizing individual drug usage habits. In [11], the microblog platform Twitter was examined as a tool for measuring public opinion of opioid use and they found that a majority of the opioid related tweets described non-medical usage, such as to obtain a 'high' or to sleep. The findings in [12] confirmed that discussion of prescription drug abuse can not only be found on twitter, but that the social circles of the corresponding participants also engage in prescription drug abuse discussions. Finally, in [13] the possibility of using social media information to construct complex behavioral models was explored. Linguistic and network based inference models were developed to predict whether a user who has posted in subreddits related to the medical condition "depression" would subsequently post in the subreddit associated with "suicide ideation", indicating thereby increased severity in his or her behavioral pathology.

Unlike simple communicable diseases, whose transmission and epidemiology involves well-understood mechanisms of infection spread, substance use is a significantly more complex process spanning psychiatric, neurological, and social factors and involves multiple stages of progression. The vast majority of research in this area has focused on extracting substance use information and identifying trends. Such information, while of undoubted value in quantifying the drug epidemic at a population level, cannot be used to characterize individual behaviors and identify individual-level interventions. The techniques proposed in this paper (*vide infra*) allow precisely the creation of such individual level predictive models.

III. Method

In this section we describe the proposed method beginning with the preliminary step of data collection followed by the lexicon embedding-expansion step, subreddit discovery, and finally creation of models for user activity prediction.

A. Data Collection

Reddit provides an open ReST API [17] that allows rate-limited access to data about its subreddits, users and user generated content. To sample the data, we began with a set of 24 well known recreational drug use and addiction recovery forums. Recent posts from these forums were extracted and provided us with information from 24,551 Reddit users. For each of these users, we queried the most recent subreddit posts up to the platform imposed limit of 1,000 posts. This amounted to a dataset of 304,338 unique Reddit posts each from a user who has posted in at least once in recreational drug use or addiction recovery subreddit.

B. Capturing and Representing the Dynamic Drug Lexicon

In social media analysis, static lists of drug terms rapidly deprecate. Consider the following exchanges posted in the year 2016 as an illustration: (1) "*I now, or used to, smoke upwards of 3, 4, 5 doobies a day*", and (2) "*Friday involved a bottle of brandy, 6 pills of methylphenidate (27mg) and a few zoots*". The terms "doobies" and "zoots" refer to weed and were never encountered in previous years. The number of unique drug terms we identified for the years 2013, 2014, 2015 and 2016 were 249, 343, 448 and 478 respectively. Accordingly, a static drug list for any year becomes inadequate in subsequent years.

We propose an approach called Density Based Lexicon Expansion with Seeding (DBLES) which combines distributed vector-space word representations with seeded density-based clustering for generating a dynamic lexicon. DBLES is applied in two steps, first a word-embedding space Ω is constructed across all documents. A set of seed terms known to refer to drugs is next selected and iteratively expanded by including the nearest terms in Ω. The process converges when a stopping criterion is met which bounds the maximum pairwise distance between terms.

As a long-form text-based social platform, Reddit presents terms of interest in a variety of contexts. Therefore, we can algorithmically compute semantic relationships between terms by modeling and comparing their contexts. Towards this, we apply the Skip-Gram (SG) model for computing the vector word representations [1]. SG is a single layer neural network model where an input word is used to predict the words in the

surrounding context. To begin, each term in the input vocabulary is initialized with a random vector of d dimensions. For each term in a given input sentence, the corresponding term-vector is used to predict the surrounding words within a window c. Predictions are made with a weight matrix in the hidden layer and a log-linear regression model (softmax). The internal weights and term-vectors are optimized with backpropagation to maximize the conditional probability of the context terms. With this formulation we arrive at a set of term-vectors that best represent the semantic contexts in which the terms were found. Taken together, the set of term vectors constitutes a d-dimensional representation space which we call the *Social Semantic Term Space* (SSTS). We apply this method with the Gensim python package [14] and its implementation of the Word2Vec algorithm [1]. In Figure 1, we present a 2-dimensional projection of the SSTS based on our data. For this figure, dimensionality reduction was performed using manifold learning with the t-SNE algorithm implemented in sklearn [20]. Dimensionality reduction is achieved by minimizing the Kullback-Leibler divergence between the original high-dimensional data and its 2D embedding with respect to location of the data points. Figure 1 shows two dense regions consisting of drug and recovery terms as well as the topology of the dispersion of these terms as we move away from the corresponding cores.

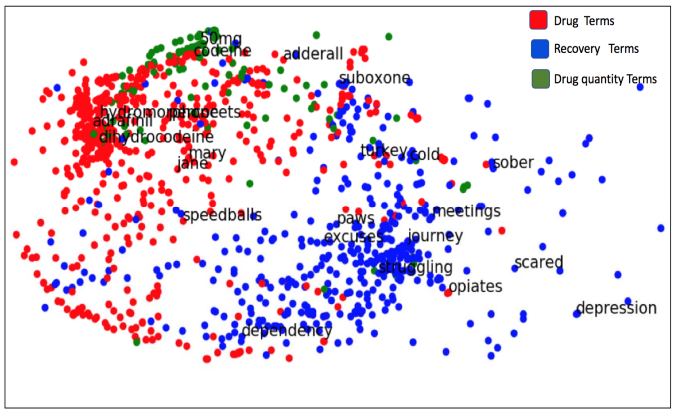

Figure 1. A two dimensional representation of the social semantic term space.

A given SSTS, Ω, contains semantically distributed term-vectors, where like-terms have greater similarity than unlike-terms. This characteristic is exploited in DBLES through a parameterized topology-based expansion. Specifically, we establish two thresholds: *neighbor_sim*, representing the minimum similarity for two terms to be considered neighbors, and *lexical_sim*, the minimum similarity between two terms in an expanded term set. We use cosine similarity here. These thresholds are selected such that *neighborl_sim* ≥ *lexical_sim*. The set expansion in DBLES is carried out as follows: we begin by initializing a lexical set L_i with a seed term from S. For each term in L_i, we expand this set by adding to it any terms from Ω that have a similarity greater than *neighbor_sim* to any term in L_i. This step is repeated as long as no two terms in L_i have a similarity less than *lexical_sim*. This process is iterated over each term in S. We end up with $|S|$ (possibly intersecting) sets of terms, $L_{1\text{-}|S|}$. The union of these sets comprises the resulting lexicon. In Table 1 we provide examples of lexicons obtained using DBLES.

C. Drug Related Subreddit Discovery with Biclustering

Like drug language, community structures also evolve on social media. For instance, the subreddit 'r/modafinil_talk', centered on the topic of eugeroic *modafinil* came into existence in the year 2014. Clustering provides an intuitive approach for extracting dynamic drug related user communities. Towards this aim, we leverage two lexicons called L_{drug} and $L_{recovery}$ generated with DBLES. The first lexicon is generated by seeding DBLES with known drug related terms and the second is generated by seeding DBLES with recovery terms.

Let $D = \{d_1,...d_n\}$ be the set of subreddits and $L = \{t_1,...t_m\}$ is the set of terms in the lexicon $\{L_{drug} \cup L_{recovery}\}$. We represent each subreddit as a term frequency vector with terms weighted by the log normalized inverse document frequency, TF-IDF [18]. The resulting $|L|$ x $|D|$ matrix M is amenable to biclustering. Unlike classical clustering, biclustering seeks to simultaneously cluster rows and columns in the matrix resulting in a set of clusters $C = \{c_1,...,c_n\}$ where $c_i = \{T \subseteq L, S \subseteq D\}$. Each cluster c_i is characterized by a set of terms and a set of subreddits. To compute this clustering, we apply spectral biclustering [2]. Here, an Eigen-decomposition of the data matrix M is performed using Singular Value Decomposition (SVD) where M is factored into three matrices $M=U\Lambda V^T$. Λ is a diagonal matrix whose values are the square root of the eigenvalues of M and the columns of U and V are the eigenvectors of M. Biclustering is then carried out by finding piecewise constant eigenvectors u and v that have the same eigenvalue. To identify drug and recovery related subreddits, we used the biclustering results to assist in manual labeling.

Seed Term	***Top five terms obtained after seed expansion***
weed	pot, marihuana, cigarettes, alcohol, cannabis
heroin	meth, opiates, cocaine, dope, coke
mdma	lsd, dxm, ketamine, molly, acid
addiction	alcoholism, addictions, abuse, habit, depression
recovery	sobriety, therapy, AA, program, treatment
sobriety	recovery, life, journey, alcoholism, career

Table 1: Terms from seed expansion. Shows examples of terms in the L_{drug} and $L_{recovery}$ lexicons based on corresponding seeding.

D. Addiction Propensity Prediction using Subreddit Activity

Our ultimate goal is to develop a model to predict whether a user will post in a recovery subreddit based on their individual (non-recovery) subreddit activity thus letting us identify users with a propensity for recovery.

We model this problem as follows: let D be the set of all subreddits, $D_{addiction}$ denote the set of addiction recovery related subreddits, and D' be the complement of $D_{addiction}$ with

respect to D, that is, $D' = \{d \in D \mid d \notin D_{addiction}\}$. We represent each user as a binary feature vector of length $|D'|$ where each element represents whether the user has posted in the corresponding subreddit. This representation allows us to apply classification methods to predict whether the user has posted to a forum in $D_{addiction}$. The performance of various classification methods investigated in the next section.

IV. Experiments, Evaluations and Results

A. Lexicon Expansion

To measure the performance DBLSs, we considered two criteria: resulting set size, and precision of results. *True positive* (TP) terms were defined to be new terms that correctly refer to drugs while *false positives* (FP) were terms that did not. Truth assignation of the resulting terms was determined through manual inspection. A critical parameter for lexical expansion is *lexical_sim* which controls minimum allowed similarity between terms in lexical sets. Its effect is investigated in Figure 2. Intuitively, as we decreased the value of *lexical_sim*, lexicon size increased, but precision remained relatively consistent in the range: $0.9 \geq lexical\_sim \geq 0.1$ with a sharp drop between 0.1 and 0. In interpreting this data, the reader may note that the value of cosine similarity is bounded between 1 and -1.

Our evaluation method only considered a term to be a *TP* if it specifically referred to a drug, however many of the FPs could be drug related at the higher values of *lexical_sim*. For example, of the seven false positives returned with the *lexical_sim* parameter set to 0.6, four described methods of ingestion, such as *sublingual, intranasal, powdered,* and *oral.* The remaining three described methods of action such as *agonist* and *inhibitor*. At $lexical\_sim \leq 0.2$, unrelated words began to appear, such as *quinoa, netbook, almond*, and *weightlifting*.

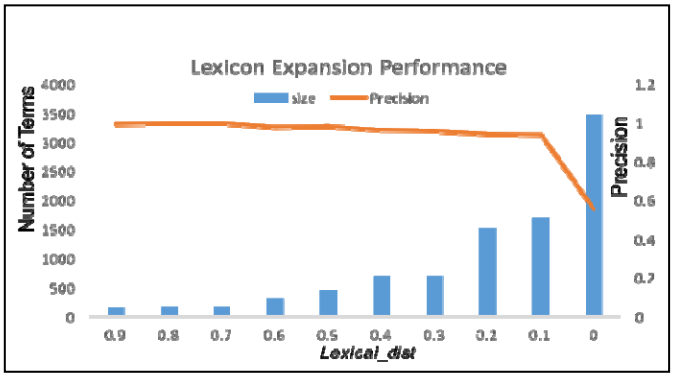

Figure 2: Analysis of Lexicon Expansion. The figure shows the effect of the *lexical_sim* parameter on the lexicon expansion. Not surprisingly, the lexicon size increases as we loosen the similarity requirements. Precision, however, does not drop off significantly until *lexical_sim* = .1

B. Drug Related Subreddit Discovery and Characterization

Our biclustering approach allows for ease of manual categorization by presenting subreddits with similar content for labeling. As the majority of the 16,208 subreddits in our data set were not drug related, biclustering correctly grouped a large majority of the subreddits into one cluster of 11,677 subreddits. The remaining clusters were predominantly drug-related or recovery-related. Table 2 shows several representative subreddit-term clusters. These clusters were used to aid manual categorization of subreddits as "Recovery Related", "Drug Related", or "Other", resulting in identification of 117 drug-related forums and 29 recovery-related forums.

Such a partitioning allows analysis of user behavior based on forum activity. Especially of interest to us is the ability to predict activity in an addiction recovery forum based on activity in other specific forums. To do so, for each non-recovery subreddit *s*, we computed the probability that a user will post in any recovery forum *r* given that they have posted in *s*. These results are presented in Figure 3. Specifically, Figure 3.1 summarizes these results for explicit drug related forums and Figure 3.2 summarizes them for non-drug use related forums. Recall that the sampling method used to generate the data guarantees that all users have posted in a drug related forum, so the charts in Figure 3.2 show the probability that a user has posted in an addiction recovery forum given that they have posted in *both* a drug related forum and the current non drug related forum. The joint probability in the condition helps explains the high probabilities shown. Interestingly, the *opiates* subreddit showed the highest value of drug–specific forums which agrees with the current trend of increasing opiate addiction. For the non-drug specific subreddits, four of the top five covered topics of relationships, the next most representative category discussed mental health, followed by legal/financial advice.

Subreddits	Terms
leaves	*Battle, program, family, addiction ,career, misery, struggle, crutch, recovery, abuse*
Depression, Trees	*Addictive, dysfunction, marijuana, pot, dysfunction, prozac, psychosis, brain*
Personalfincance, AtheistTwelveSteppers, Advice, Relationshipadvice, Divorce, legaladvice, Meditation	*Cocaine, mirtazapine, weed, Zoloft, Lexapro, Effexor*
Addiction	*Habbit, opiate, epidemic, opioids,ir, strips*

Table 2: Representative Subreddit - Term Clusters. Representative results from the biclustering method are shown. The subreddit *leaves* is focused on marijuana addiction recovery, its associated terms paint a vivid picture of the conversations present in the subreddit.

C. Addiction Propensity Prediction

We used user activity as a predictor of addiction propensity. The ground truth label for each user is determined by whether or not they have posted in drug recovery related forums. Our dataset comprised of 24,551 users, 8,697 of whom posted in recovery forums. Results are summarized in Table 3. The *k*-Nearest Neighbors (*k*-NN) classifier with k=11 had the best overall performance with an F1 score of 0.848. However, Random Forests showed the best precision with 0.914 and *k*-NN with k=3 had the best recall at 0.843. Different values of k = 5, 7, 9, and 13 were also tried. Their

	Performance on test data		
Method	Precision	Recall	F1
K-NN (3)	792	**.843**	.817
K-NN (11)	.900	**.803**	**.848**
Random forests	**.914**	.782	.843
Logistic Regression	.685	.692	.689
Naïve Bayes	.702	.382	.495

Table 3: **Evaluation of addiction propensity prediction performance.** Random Forests and K-Nearest Neighbors show the highest precision and recall respectively. The best F1 score is obtained using *k*-Nearest Neighbors with the value of *k*=11.

precision and recall values were found to range between the values for *k*=3 and *k*=11. Therefore, only results for *k*=3 and 11 are showed in Table 3.

Classification algorithms can also provide us with insight into the predictive power of individual features, in this case subreddit activity. Accordingly, we observed the importance of several features in predicting addiction propensity in Fig. 3.

	Performance		
Method	Precision	Recall	F1
K-NN (3-NN)	.611 ± .034	.797 ± .058	.690 ± .033
K-NN (11-NN)	.644 ± .034	**.863 ± .032**	**.737 ± .028**
Random forests	**.739 ± .047**	.679 ± .077	**.706 ± .057**

Table 4: Addition Propensity Prediction on Twitter. Performance evaluations shown are with 10-fold crossvalidation.

For the Random Forests classifier, importance is defined as the average Gini importance values of each feature for each classifier in the ensemble [15]. Here, we do not see opiates as the highest ranked subreddits; eight subreddits have more predictive importance, including the general subreddit *drugs*, and the marijuana focused subreddit *trees*.

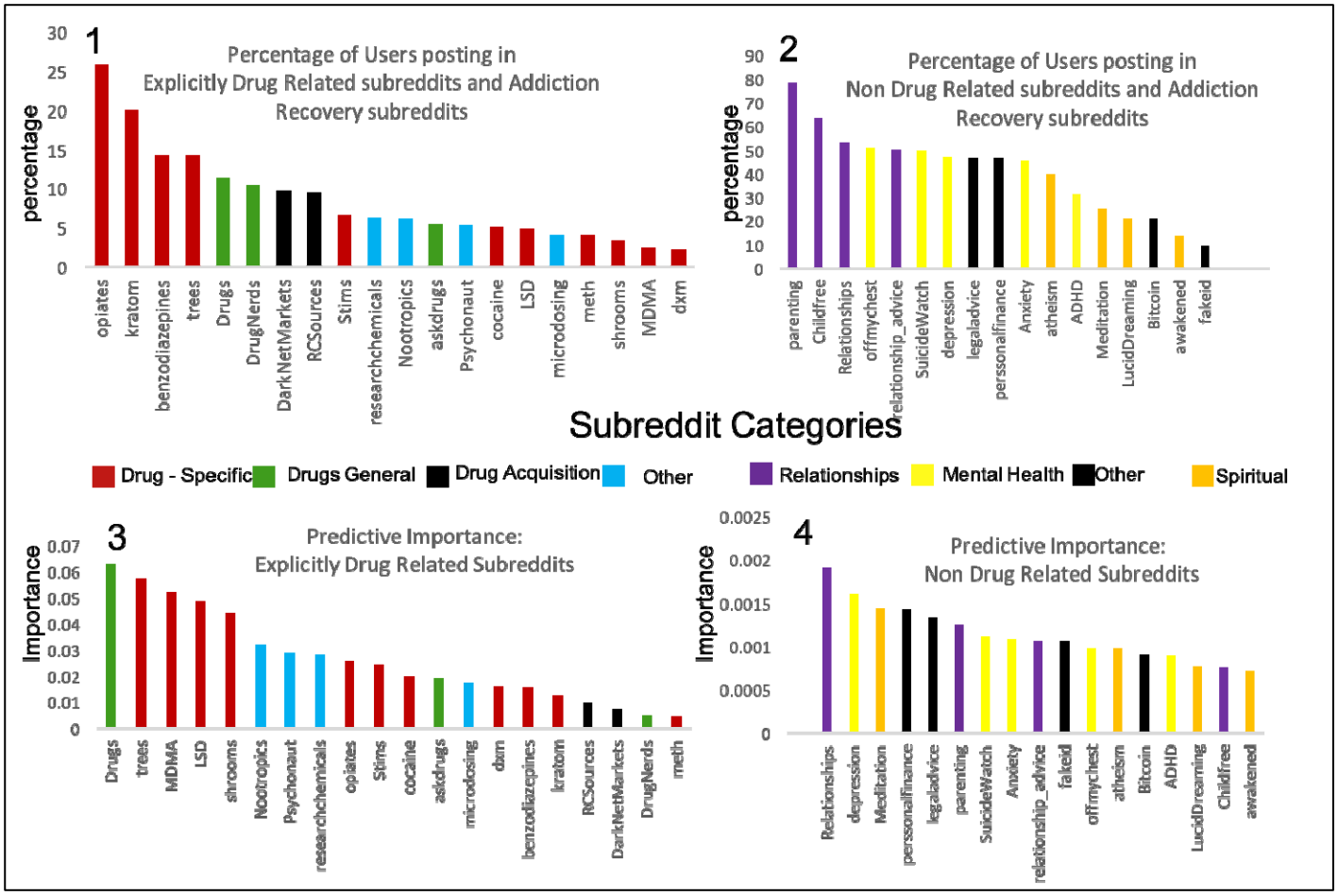

Figure 3: Subreddit Characterizations: 3.1 shows the percentage of users who posted in the explicitly drug related subreddits (as shown) and in an addiction recovery subreddit (not shown). 25.8% of users who posted in the *opiates* subreddit also posted in a recovery subreddit. **3.2** shows the percentage of users who posted in an explicitly drug related subreddit, the non drug related subreddits shown and an addiction recovery subreddit. Interesting categories emerge, such as subreddits centered around relationships and mental health issues. **3.3** Shows the predictive importance of the different subreddits based on the Gini-importance values computed during the Random Forests classification. The sum of the Gini-importance values across all 16,208 subreddits will equal 1. The subreddit *opiates* does not have the highest importance, this may be explained by the larger user communities in the subreddits with higher importance. **3.4** Shows the predictive importance of the highest ranked non-drug related subreddits.

D. Addiction Propensity Prediction using Twitter data

We adapted our method to a Twitter dataset comprised of 1401 Twitter users, 673 of whom, directly mention their sobriety from drug and alcohol use and 728 explicitly discuss their current drug use. For the feature space, we substituted subreddit activity for hashtag usage. The results reported in Table 4 suggest that our approach generalizes across domains.

V. Conclusions

We have developed an algorithmic approach to address three fundamental challenges in detecting addiction trends using social media data. Our method helps characterize the discussions about drugs, the social media forums that are used for such discussions, and how drug user activity can be related to the propensity for drug addiction and recovery. In particular, our approach constitutes a powerful tool in focusing recovery efforts on drug users who are most amenable. In so doing, we have also determined intriguing patterns, such as the high percentage of users in the "opiates" community that interact with the "recovery" communities, as well as some latent relationships such as those between the "parenting" and "recovery" communities. These results open the possibility of developing deeper understanding of the complex factors that drive and influence addiction. Finally, the method presented in this paper represents a fundamental advancement in the area and can be expected to find broad applicability in the design of effective interventions for arresting the substance-use epidemic.

References

[1] T. Mikolov, et al., "Distributed representations of words and phrases and their compositionality", Advances in neural information processing systems. (2013).

[2] Y. Kluger, et al., "Spectral biclustering of microarray data: coclustering genes and conditions", Genome research 13.4 (2003): 703-716.

[3] L. Paulozzi, Y. Xi, "Recent changes in drug poisoning mortality in the United States by urban-rural status and by drug type", Pharmacoepidemiology and Drug Safety, 17 (2008): 997–1005.

[4] R. A. Rudd, et al, "Increases in drug and opioid overdose deaths-United States, 2000-2014", MMWR: Morbidity and mortality weekly report 64 (2016) : 1378-1382.

[5] F. Schifano, A. Ricciardi, O. Corazza, P. Deluca, Z. Davey, C. Rafanelli, "Psychonaut web mapping: new drugs of abuse on the web: the role of the Psychonaut web mapping project", Riv Psichiatr, 45 (2010) : 88–93.

[6] P. Miller, A. Sonderlund, "Using the internet to research hidden populations", Addiction 105 (2010) : 1557–67.

[7] E. Boyer, M. Shannon, P. Hibberd, "The internet and psychoactive substance use among innovative drug users", Pediatrics 115 (2005) : 302–5.

[8] E. Boyer, M. Shannon, P. Hibbert, "Web sites with misinformation about illicit drugs", New England Journal of Medicine 345 (2001) : 469–71.

[9] F. Schifano, P. Deluca, "Psychonaut 2002 research group: searching the internet for drug-related web sites: analysis of online available information on ecstasy (mdma)", American Journal of Addiction, 16 (2007) : 479–83.

[10] D. Cameron, et al, "PREDOSE: A semantic web platform for drug abuse epidemiology using social media", Journal of biomedical informatics 46 (2013) : 985-997.

[11] B. Chan, A. Lopez, U. Sarkar, "The Canary in the Coal Mine Tweets: Social Media Reveals Public Perceptions of Non-Medical Use of Opioids", PloS One, 10 (2015): e0135072.

[12] C. L. Hanson, et al, "An exploration of social circles and prescription drug abuse through Twitter", Journal of medical Internet research, 15 (2013): e189.

[13] M. De Choudhury, et al. "Discovering shifts to suicidal ideation from mental health content in social media", Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. ACM, (2016)

[14] https://radimrehurek.com/gensim/

[15] Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.

[16] http://expandedramblings.com/index.php/reddit-stats/

[17] https://www.reddit.com/dev/api/

[18] G. Salton, C. Buckley, "Term-weighting approaches in automatic text retrieval", *Information processing & management* 24.5 (1988): 513-523.

[19] President Commission on the Opioid Crisis, Interim Report, July 31, 2017(https://www.whitehouse.gov/sites/whitehouse.gov/files/ondcp/commission-interim-report.pdf)

[20] F. Pedregosa, et al. "Scikit-learn: Machine learning in Python." *Journal of Machine Learning Research* 12.Oct (2011): 2825-2830.