The Workshops of the Thirty-Second
AAAI Conference on Artificial Intelligence

Utilizing Social Media to Combat Opioid Addition Epidemic: Automatic Detection of Opioid Users from Twitter

Yiming Zhang, Yujie Fan
Department of Computer Science
and Electrical Engineering
West Virginia University, WV, USA
{ymzhang, yf0004}@mix.wvu.edu

Yanfang Ye,* Xin Li
Department of Computer Science
and Electrical Engineering
West Virginia University, WV, USA
{yanfang.ye, xin.li}@mail.wvu.edu

Erin L. Winstanley
Department of Pharmaceutical Systems
and Policy School of Pharmacy
West Virginia University, WV, USA
erin.winstanley@hsc.wvu.edu

Abstract

Opioid (e.g., heroin and morphine) addiction has become one of the largest and deadliest epidemics in the United States. To combat such deadly epidemic, in this paper, we propose a novel framework named *AutoOPU* to automatically detect the opioid users from Twitter, which will assist in sharpening our understanding toward the behavioral process of opioid addiction and treatment. In *AutoOPU*, to model the users and posted tweets as well as their rich relationships, we first introduce a heterogeneous information network (HIN) for representation. Then we use meta-structure based approach to characterize the semantic relatedness over users. Afterwards, we integrate content-based similarity and relatedness depicted by each meta-structure to formulate a similarity measure over users. Further, we aggregate different similarities using multi-kernel learning, each of which is automatically weighted by the learning algorithm to make predictions. To the best of our knowledge, this is the first work to use multi-kernel learning based on meta-structures over HIN for biomedical knowledge mining, especially in drug-addiction domain. Comprehensive experiments on real sample collections from Twitter are conducted to validate the effectiveness of our developed system *AutoOPU* in opioid user detection by comparisons with other alternative methods.

1. Introduction

Opioids (NIDA 2015) are a group of drugs which include the illegal drug heroin and powerful pain relievers by legal prescription (e.g., morphine, oxycodone). Opioid addiction has become one of the largest and deadliest epidemics in the U.S. (Murthy 2016). There was a skyrocketing increase of opioid related death in the past decade: according to National Institute on Drug Abuse, in 2014, 18,893 Americans died from opioid analgesics and 10,574 people died from heroin overdose, both reflecting significant increase from 2001 (NIDA 2015). Opioid addiction has also turned into a serious global concern because of its negative health, social and economic impacts (e.g., family breakdown, domestic violence, child abuse). Opioid addiction is a chronic mental illness that requires long-term treatment and care. It is a psychiatric challenge because of high relapse and drop-out rates. Although Medication Assisted Treatment (MAT) using methadone or buprenorphine has been proven to provide best outcomes for opioid addiction recovery, stigma (i.e., bias) associated with MAT has limited its utilization. Therefore, there is an imminent need for novel tools and methodologies to gain new insights into the behavioral processes of opioid addiction and treatment.

In recent years, the role of social media in biomedical knowledge mining, such as drug pharmacology and interactive healthcare, has become more and more important. Due to the increasing use of the Internet, never-ending growth of data are generated from the social media which offers opportunities for the users to freely share opinions and experiences in online communities. For example, Twitter, as one of the most popular social media platforms, has more than 140 million active users posting over 500 million 140 character tweets every day (LiveStats 2013). A larger number of Twitter users are willing to share their experiences of using opioids (e.g., "*I have a crippling heroin addiction and it is destroying my life*"), and perceptions toward MAT (e.g., "*heroin; I think this model of treatment (methadone) needs to be made available in the US, as it is the most effective treatment for opioid.*"). Therefore, the data from social media may contribute information beyond the knowledge of domain professionals (e.g., psychiatrists and epidemics researchers) and could potentially assist in sharpening our understanding toward the behavioral process of opioid addiction and treatment.

To combat the opioid addiction epidemic and promote the practice of MAT, in this paper, we propose a novel framework named *AutoOPU*, a multi-kernel learning model based on meta-structures over heterogeneous information network (HIN), to automatically detect the opioid users from Twitter. To model the users and posted tweets as well as their rich semantic relationships, in *AutoOPU*, we first introduce a HIN (Han et al. 2010; Sun and Han 2012) for representation, which is capable to be composed of different types of entities and relations. To capture the complex relationship (e.g., two users are relevant if they have posted tweets which are talked publicly to the same person, and have also discussed the same topic), we use a meta-structure (Huang et al. 2016) based approach to characterize the semantic relatedness over users. Then, we further integrate content-based similarity (i.e., similarity of users' posted tweets) and relat-

*Corresponding author.

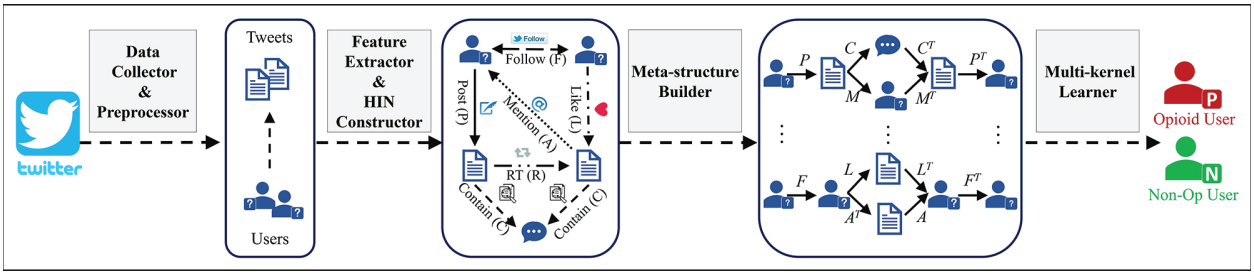

Figure 1: System architecture of *AutoOPU*.

edness depicted by each meta-structure to formulate a similarity measure over users. Later, we aggregate different similarities using multi-kernel learning (Sun et al. 2010), each of which is automatically weighted by the learning algorithm to make predictions. The major contributions of our work can be summerized as follows:

- This is a *pioneer work* to automatically detect opioid users from Twitter for the study of opioid epidemic; the proposed framework is also extendable to the surveillance analysis through social media for other drugs of interests.
- We propose *novel feature representation and user relatedness characterization* to describe Twitter users. Based on different kinds of relationships (i.e., user-user, user-tweet, tweet-tweet, tweet-topic relations) through different types of entities (i.e., user, tweet, topic), the users will be represented by a HIN, and a meta-structure based approach will be used to characterize the relatedness between users. To utilize both content- and relation-based information, we integrate similarity of users' posted tweets and relatedness depicted by each meta-structure to formulate a similarity measure over users. The proposed solution provides a more convenient way to express the complex relationships in social network than traditional approaches.
- We present a *multi-kernel learner to aggregate different similarities* defined by different meta-structures combined with content-based information. This is a very natural way to aggregate different similarities formulated by meta-structures but to our best knowledge is a first attempt.
- We develop *a practical system AutoOPU* integrated with the proposed method for automatic opioid user detection, based on a large-scale data collection from Twitter and manually constructed ground-truth labels. Comprehensive experimental studies are conducted to validate the effectiveness of our developed system in comparisons with other alternative approaches.

2. System Architecture

Figure 1 shows the system architecture of *AutoOPU*, which is developed to automatically detect opioid users from Twitter. It consists of the following five major components:

- ***Data Collector and Preprocessor.*** We first develop the web crawling tools to collect the tweets including opioid-related keywords (e.g. *heroin*, *morphine*, street names or slangs like *black tar*) as well as users' profiles from Twitter. To protect the users' privacy, we use UserID to represent each individual user whose information is kept anonymous. For the collected tweets, the preprocessor will further remove all the links, punctuation and stopwords, and conduct lemmatization using Stanford CoreNLP (Manning et al. 2014).
- ***Feature Extractor and HIN Constructor.*** A bag-of-words (Yang et al. 2007) feature vector will be extracted to represent each user's posted tweet(s). Besides, the relationships among users, tweets and topics will be further analyzed, such as, i) *user-follow-user* (i.e., two users follow each other), ii) *user-like-tweet* (i.e., one user shows appreciation for a posted tweet), iii) *tweet-mention-user* (i.e., one user is @ (i.e. mentioned/talked publicly) in the tweet), iv) *tweet-RT-tweet* (i.e., a repost of a tweet), and v) *tweet-contain-topic* (i.e. a tweet contains a specific topic). Based on the extracted features, a structural HIN is then constructed. (See Section 3.1 for details.)
- ***Meta-structure Builder.*** In this module, different meta-structures are first built from HIN to capture the relatedness between users. Then, we integrate similarity of users' posted tweets and relatedness depicted by each meta-structure to formulate a set of similarity measures over users. (See Section 3.2 for details.)
- ***Multi-kernel Learner.*** Given the similarity matrices over users defined by different meta-structures combined with content-based information constructed by the previous component, a multi-kernel learner which treats each matrix as a kernel, is used to weight the importance of each similarity. Then, a more powerful kernel is generated through the aggregation of these similarities for automatic opioid user detection. (See Section 3.3 for details.)
- ***Opioid User Detector.*** For each unlabeled user, his/her posted tweets and the above-mentioned relationships will be extracted; using the constructed classification model, the user will then be labeled as either opioid user or not.

3. Proposed Method

In this section, we introduce the detailed approaches of how we represent Twitter users, and how we solve the problem of opioid user detection based on this representation.

3.1 HIN Construction

As the above discussion, to detect opioid users from Twitter, we not only utilized users' posted tweets but also the rich se-

mantic relationships among the users and posted tweets. To characterize the relatedness of two users, we consider various kinds of relationships which include the followings.

- ***R1***: To describe the relation of a user and his/her posted tweet, we generate the ***user-post-tweet*** matrix $\mathbf{P}$ where each element $p_{i,j} \in \{0, 1\}$ denotes if user *i* posts tweet *j*.
- ***R2***: To denote the relation that a user appreciates a tweet, we generate the ***user-like-tweet*** matrix $\mathbf{L}$ where each element $l_{i,j} \in \{0, 1\}$ means if user *i* likes tweet *j*.
- ***R3***: If two users follow each other (i.e., called *tweeps*), it could imply that they might be friends or have similar interests. To represent such user-user relationship, we generate the ***user-follow-user*** matrix $\mathbf{F}$ where each element $f_{i,j} \in \{0, 1\}$ denotes if user *i* and user *j* follow each other.
- ***R4***: Like in the physical world, users can talk publicly to another in Twitter: if a tweet includes the symbol of @ followed by a user name, it means that the user is mentioned and talked publicly in this tweet. To describe this type of tweet-user relationship, we build the ***tweet-mention-user*** matrix $\mathbf{A}$ where each element $a_{i,j} \in \{0, 1\}$ indicates if tweet *i* mentions user *j*.
- ***R5***: A tweet can be a repost of another tweet. To represent such relationship between two tweets, we build the ***tweet-RT-tweet*** matrix $\mathbf{X}$ where element $x_{i,j} \in \{0, 1\}$ denotes if tweet *i* or tweet *j* is a repost of the other.
- ***R6***: To represent the relation that a tweet contains a specific topic, we generate the ***tweet-contain-topic*** matrix $\mathbf{C}$ where each element $c_{i,j} \in \{0, 1\}$ indicates if tweet *i* contains topic *j*. In our application, we use Latent Dirichlet allocation (Blei et al. 2003) for the topic extraction from the posted tweets.

In order to depict users, tweets, topics and the rich relationships among them, it is important to model them in a proper way so that different kinds of relations can be better and easier handled. We introduce how to use HIN, which is capable to be composed of different types of entities and relations, to represent the users by using the features described above. We first present some concepts related to HIN.

Definition 1 ***Heterogeneous information network (HIN)*** *(Sun and Han 2012). A HIN is defined as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with an entity type mapping ϕ: $\mathcal{V} \rightarrow \mathcal{A}$ and a relation type mapping ψ: $\mathcal{E} \rightarrow \mathcal{R}$, where $\mathcal{V}$ denotes the entity set and $\mathcal{E}$ is the relation set, $\mathcal{A}$ denotes the entity type set and $\mathcal{R}$ is the relation type set, and the number of entity types $|\mathcal{A}| > 1$ or the number of relation types $|\mathcal{R}| > 1$.*

Definition 2 ***Network schema*** *(Sun et al. 2011b). The network schema for a HIN $\mathcal{G}$, denoted as $\mathcal{T}_{\mathcal{G}} = (\mathcal{A}, \mathcal{R})$, is a graph with nodes as entity types from $\mathcal{A}$ and edges as relation types from $\mathcal{R}$.*

HIN not only provides the network structure of data associations, but also a high-level abstraction of the categorical association. Based on the definitions above, the network schema for HIN in our application is shown in Figure 2.

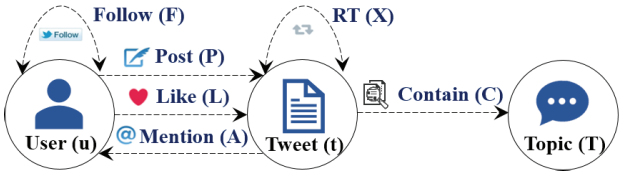

Figure 2: Network schema for HIN.

3.2 Meta-structure Based Relatedness

The different types of entities and different relations between them motivate us to use a machine-readable representation to enrich the semantics of relatedness among users. Meta-path (Sun et al. 2011b) is used in HIN to formulate the semantics of higher-order relationships among entities.

Definition 3 ***Meta-path*** *(Sun et al. 2011b) . A meta-path $\mathcal{P}$ is a path defined on the graph of network schema $\mathcal{T}_{\mathcal{G}} = (\mathcal{A}, \mathcal{R})$, and is denoted in the form of $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \ldots \xrightarrow{R_L} A_{L+1}$, which defines a composite relation $R = R_1 \cdot R_2 \cdot \ldots \cdot R_L$ between types A_1 and A_{L+1}, where $\cdot$ is relation composition operator, and L is the length of $\mathcal{P}$.*

An example of a meta-path for users based on HIN schema shown in Figure 2 is: $user \xrightarrow{post} tweet \xrightarrow{contain} topic \xrightarrow{contain^{-1}} tweet \xrightarrow{post^{-1}} user$, which states that two users can be connected through their posted tweets containing same topics; another example is $user \xrightarrow{post} tweet \xrightarrow{mention} user \xrightarrow{mention^{-1}} tweet \xrightarrow{post^{-1}} user$, which denotes that two users are related by their posted tweets mentioning same users. Although meta-path has been shown to be useful for relatedness measure between users (Sun et al. 2011b; Luo et al. 2014), it fails to capture a more complex relationship, e.g., two users have posted tweets discussed the same topic and have also talked publicly to (i.e., mentioned) the same person. This calls for a better characterization to handle such complex relationship. Meta-structure (Huang et al. 2016) is proposed to use a directed acyclic graph of entity and relation types to capture more complex relationship between two HIN entities. The concept of meta-structure is given as following (Huang et al. 2016):

Definition 4 ***Meta-structure*** *(Huang et al. 2016). A meta-structure $\mathcal{S}$ is a directed acyclic graph with a single source node n_s and a single target node n_t, defined on a HIN schema $\mathcal{T}_{\mathcal{G}} = (\mathcal{A}, \mathcal{R})$. Formally, $\mathcal{S} = (N, M, n_s, n_t)$, where N is a set of nodes and M is a set of edges. For any node $x \in N, x \in \mathcal{A}$; for any link $(x, y) \in M, (x, y) \in \mathcal{R}$.*

In our application, based on the HIN schema displayed in Figure 2, we generate five meaningful meta-structures to characterize the relatedness over users (i.e., ***SID1–SID5*** shown in Figure 3: *left*). For example, ***SID1*** depicts that two users are related if they have posted tweets discussed same topics and have also talked publicly to (i.e., mentioned) same people; while ***SID4*** describes that two users are connected if the tweets they like have discussed same topics and have also reposted same tweets from other people. Actually, a

meta-path is a special case of a meta-structure (e.g., *PID1* and *PID2* are particular cases of *SID1*). In Figure 3, the meta-paths of *PID1–PID8* (*right*) are the special cases of the constructed meta-structures *SID1–SID5* (*left*). But meta-structure is capable to express more complex relationship in a convenient way.

To compute the relatedness over users using a particular meta-structure designed above, we use the following commuting matrix to give a general form.

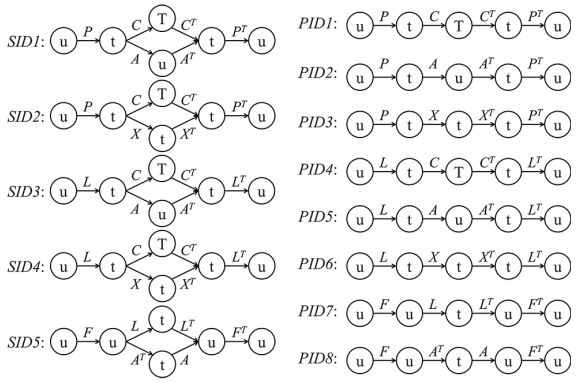

Figure 3: Meta-structures (*left*) and meta-paths (*right*). (The symbols in the circles are abbreviations shown in Figure 2.)

Definition 5 ***Commuting matrix*** *(Huang et al. 2016). Given a network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and its network schema $\mathcal{T}_\mathcal{G} = (\mathcal{A}, \mathcal{R})$, a commuting matrix $\boldsymbol{M}_\mathcal{S}$ for the meta-structure $\mathcal{S} = A_1 \rightarrow A_2 \overset{\nearrow A_3 \searrow}{\searrow A_4 \nearrow} A_5 \rightarrow A_6$ is defined as $\boldsymbol{M}_\mathcal{S} = \boldsymbol{W}_{A_1A_2}[(\boldsymbol{W}_{A_2A_3}\boldsymbol{W}_{A_3A_5}) \circ (\boldsymbol{W}_{A_2A_4}\boldsymbol{W}_{A_4A_5})]\boldsymbol{W}_{A_5A_6}$, where $\boldsymbol{W}_{A_iA_j}$ is the adjacency matrix between types A_i and A_j, $\circ$ denotes the Hadamard product (Horn 1990) of two matrices, whose computation complexity is $O(n^2)$, n is the number of rows of the matrix $\boldsymbol{W}_{A_2A_3}$. $\boldsymbol{M}_\mathcal{S}(i,j)$ represents the number of structure instances between entity $x_i \in A_1$ and entity $y_j \in A_6$ under meta-structure $\mathcal{S}$.*

For example, the commuting matrix of users computed using *SID1* is $\mathbf{P}[(\mathbf{C}\mathbf{C}^T) \circ (\mathbf{A}\mathbf{A}^T)]\mathbf{P}^T$, whose element $\mathbf{M}_{\mathcal{S}_1}(i,j)$ denotes the number of tweet pairs posted by user i and user j which contain same topics and also mention same people. Table 1 shows the commuting matrix of each meta-structure and the description of its element.

After characterizing the relatedness of users, we utilize both content- and relation-based information to measure the similarity over users: we integrate similarity of users' posted tweets and relatedness depicted by meta-structure to form a similarity measure matrix over users. The similarity matrix over users is denoted as $\mathbf{Q}$, whose element is a combination of content-based similarity and meta-structure based relatedness. We define similarity matrix $\mathbf{Q}_{\mathcal{S}_k}$ based on $\mathbf{M}_{\mathcal{S}_k}$ as:

$$\mathbf{Q}_{\mathcal{S}_k}(i,j) = [1 + log(\mathbf{M}_{\mathcal{S}_k}(i,j) + 1)] \times tSim(i,j), \quad (1)$$

where $\mathbf{M}_{\mathcal{S}_k}(i,j)$ is the relatedness between user *i* and *j* under meta-structure $\mathcal{S}_k$, $tSim(i,j)$ is the similarity between two users' posted tweets. A user may post multiple tweets including opioid-related keywords. Thus, for each user, we convert his/her posted tweet(s) into a bag-of-words feature vector and use cosine similarity measure (Mihalcea et al. 2006) to estimate the closeness of two users' posted content.

3.3 Multi-Kernel Learning

Different meta-structures capture the relatedness over users at different views, i.e., *SID1–SID5*. Since HIN can naturally provide us different relatedness with different semantics, instead of using a single meta-structure to depict the relatedness between users, we propose to use a multi-kernel learning algorithm to automatically incorporate different similarities based on different meta-structures and weight each of them for user classification.

Supposed that there are *K* meta-structures $\mathcal{S}_k$ ($k = 1, 2, ..., K$), we can calculate their corresponding commuting matrices $\mathbf{M}_{\mathcal{S}_k}$ ($k = 1, 2, ..., K$). Then, we use Eq.(1) to compute the similarity matrix $\mathbf{Q}_{\mathcal{S}_k}$ ($k = 1, 2, ..., K$) based on $\mathbf{M}_{\mathcal{S}_k}$. We treat each similarity matrix $\mathbf{Q}_{\mathcal{S}_k}$ as a kernel in multi-kernel learning model. If the matrix $\mathbf{Q}_{\mathcal{S}_k}$ is not a kernel (not a positive semi-definite matrix), we simply use the trick to remove the negative eigenvalues. A new kernel is formed using the linear combination of the computed kernels, which can be defined as (Sun et al. 2010; Gönen and Alpaydın 2011):

$$\mathbf{Q}' = \sum_{k=1}^{K} \gamma_k \mathbf{Q}_{\mathcal{S}_k}, \quad (2)$$

where the weights $\gamma_k \geq 0$ and satisfy $\sum_{k=1}^{K} \gamma_k = 1$.

To learn the weight of each kernel, we assume we have a set of labeled data $\{\mathbf{x}_i, y_i\}_{i=1}^N$, where $\mathbf{x}_i$ is the *i*-th user, $y_i \in \{+1, -1\}$ is the corresponding label (+1 denotes opioid user while -1 means non-opioid user). Then we use the *p*-norm multi-kernel learning framework (Sun et al. 2010) with following objective function for parameter learning:

$$\min_{\mathbf{w}>0, \xi_i, \gamma_i \geq 0} \frac{1}{2} \sum_k \|\mathbf{w}_k\|^2 / \gamma_k + C \sum_i \xi_i + \frac{\lambda}{2} (\sum_k \gamma_k^p)^{\frac{2}{p}},$$
$$s.t. \quad y_i (\sum_k \mathbf{w}_k^T \varphi_k(\mathbf{x}_i) + b) \geq 1 - \xi_i, \quad (3)$$

where $\mathbf{w}_k$ is a weight vector associated with each kernel. For each data $\{\mathbf{x}_i, y_i\}$, the slack parameter ξ_i is introduced to allow mis-classification. $\varphi_k(\mathbf{x}_i)$ is a nonlinear mapping function of features in the Hilbert space that defines the kernel, where $\mathbf{Q}_{\mathcal{S}_k}(i,j) = \varphi_k(x_i)^T \varphi_k(x_i)$. Then by applying the representation theorem, we have $\mathbf{w}_k = \sum_i \alpha_i \varphi_k(\mathbf{x}_i)$. α_i can be solved using the dual formulation, and non-zero α_i's lead to the support vectors. For another set of parameters γ_k, the *p*-norm $(\sum_k \gamma_k^p)^{\frac{2}{p}}$ is used for regularization. Empirically, 2-norm performs best in our application and is thus applied to our problem throughout the paper. After the optimization, the weights γ_k's are obtained to reveal the importance of different similarities based on different meta-structures. For a user $\mathbf{x}$,

$$\sum_k \mathbf{w}_k \varphi_k(\mathbf{x}) + b, \quad (4)$$

is used to predict whether he/she is an opioid user. The opioid user detection procedure is given in Algorithm 1.

Table 1: The description of each meta-structure.

SID	Commuting matrix $\mathbf{M}$	Description of each element m_{ij} in $\mathbf{M}$
1	$\mathbf{P}[(\mathbf{C}\mathbf{C}^T) \circ (\mathbf{A}\mathbf{A}^T)]\mathbf{P}^T$	# of tweet pairs posted by user *i* and *j* that contain same topics and mention same people
2	$\mathbf{P}[(\mathbf{C}\mathbf{C}^T) \circ (\mathbf{X}\mathbf{X}^T)]\mathbf{P}^T$	# of tweet pairs posted by user *i* and *j* that contain same topics and repost same tweets
3	$\mathbf{L}[(\mathbf{C}\mathbf{C}^T) \circ (\mathbf{A}\mathbf{A}^T)]\mathbf{L}^T$	# of tweet pairs liked by user *i* and *j* that contain same topics and mention same people
4	$\mathbf{L}[(\mathbf{C}\mathbf{C}^T) \circ (\mathbf{X}\mathbf{X}^T)]\mathbf{L}^T$	# of tweet pairs liked by user *i* and *j* that contain same topics and repost same tweets
5	$\mathbf{F}[(\mathbf{L}\mathbf{L}^T) \circ (\mathbf{A}^T\mathbf{A})]\mathbf{F}^T$	# of tweep pairs of user *i* and *j* who like same tweets and are mentioned in same tweets

Algorithm 1 Automatic Opioid User Detection Algorithm

Input: Training dataset T_r, testing dataset T_e
Output: Labels of users in T_e

1: Generate matrix **P**, **L**, **F**, **A**, **X** and **C** for T_r;
2: Define meta-structure set $\mathbf{S}_{\mathcal{S}} = \{\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_5\}$ based on the six matrices above;
3: **for** each meta-structure $\mathcal{S}_k$ ($k = 1, 2, ..., 5$) in $\mathbf{S}_{\mathcal{S}}$ **do**
4: Compute $\mathbf{M}_{\mathcal{S}_k}$ based on Definition 5;
5: Compute $\mathbf{Q}_{\mathcal{S}_k}$ using Eq.(1);
6: **end for**
7: Let each $\mathbf{Q}_{\mathcal{S}_k}$ be a kernel in the multi-kernel learning model, and compute the weight vector $\mathbf{w}_k$ for each kernel by optimizing Eq.(3);
8: **for** each user **x** in T_e **do**
9: Predict its label using Eq.(4);
10: **end for**

4. Experimental Results And Analysis

In this section, we show two sets of experimental studies using real sample collections from Twitter to fully evaluate the performance of our developed system *AutoOPU* for automatic opioid user detection: (1) In the first set of experiments, based on HIN schema, we fully assess the performance of our proposed method; (2) In the second set of experiments, we evaluate our developed system *AutoOPU* which integrates our proposed method by comparisons with other alternative classification approaches. We use accuracy (ACC) and F1 measure, as well as Area Under the Curve (AUC) (Hou et al. 2017) as the evaluation metric.

4.1 Data Collection and Annotation

To obtain the data from Twitter, we develop web crawling tools to collect the tweets including keywords of opioids (e.g., heroin, morphine) and the common *street* or *slang* names (e.g., black tar, RMS, subutex), as well as users' profiles in a period of time. By the date, we have collected over ***4,447,507*** opioid-related tweets from nearly ***4,051,423*** users through March 2007 to January 2017.

As heroin addiction occupies the majority of today's opioid addiction, in this paper, we first study the heroin-related tweets and their related users. To obtain the prelabeled data for training, based on the collected data (including the posted tweets, users' profiles and their social relations, etc.), five groups of annotators (i.e., ***18 persons***) with knowledge from domain professional (i.e., psychiatrist) ***spent three months to label*** whether they are opioid users or not by cross-validations. The mutual agreement is above 95%, and only the ones with agreements are retained. The annotated dataset (denoted as $\mathbf{DB_a}$) consists of 2,510 users (1,208 are labeled as opioid users and 1,302 are non-opioid users) related to 20,780 tweets (11,139 are posted by opioid users and 9,641 are posted by non-opioid users) .

4.2 Evaluation of the Proposed Method

In this set of experiments, based on the annotated dataset $\mathbf{DB_a}$, we fully evaluate our proposed method: (1) based on the HIN schema (described in Section 3.1), we first evaluate the performance of meta-structure based method in opioid user detection by comparisons with meta-path based approach; (2) we then evaluate the proposed multi-kernel learning method for aggregation of different similarities based on different meta-structures. In the experiments, we randomly select 90% of the data for training, while the remaining 10% is used for testing.

We first construct five meta-structures (i.e., *SID1–SID5* shown in Figure 3: *left*) and generate the corresponding eight meta-paths (i.e., *PID1–PID8* shown in Figure 3: *right*). To measure the similarity over users, as described in Section 3.2, we integrate similarity of users' posted tweets and relatedness depicted by each meta-structure or meta-path to form a similarity measure matrix. We evaluate their performances for opioid user detection using Support Vector Machine (SVM). For each meta-structure or meta-path, the generated similarity measure matrix is used as the kernel fed to SVM. For SVM, we use LibSVM in our experiments and the penalty is empirically set to be 1,000.

The results in Table 2 show that each meta-structure does perform better than its corresponding meta-paths. For example, meta-paths of *PID1* and *PID2* are special cases of meta-structure *SID1*; but *SID1* works better than *PID1* and *PID2* in the problem of opioid user detection. The reason behind this is that meta-structure is more expressive to characterize a complex relatedness over users than meta-path. This also demonstrates that we can use meta-structure with subtle differences to significantly improve the quality of relation-based features and better express different relatedness over users in our application.

We then use all the generated similarity matrices based on different meta-structures as the kernels (i.e., $\mathbf{Q}_{\mathcal{S}_1}$–$\mathbf{Q}_{\mathcal{S}_5}$) and apply multi-kernel learning (described in Section 3.3) to construct a new kernel (i.e., *ID15*) fed to SVM. By comparisons, we also combine these different similarity matrices using their Laplacian scores (He, Cai, and Niyogi 2006) as

Table 2: Evaluation of the proposed method.

ID	Kernel	Commuting Matrix	ACC	F1
PID1	$\mathbf{Q}_{\mathcal{P}_1}$	$\mathbf{PCC}^T\mathbf{P}^T$	0.806	0.792
PID2	$\mathbf{Q}_{\mathcal{P}_2}$	$\mathbf{PAA}^T\mathbf{P}^T$	0.773	0.768
PID3	$\mathbf{Q}_{\mathcal{P}_3}$	$\mathbf{PXX}^T\mathbf{P}^T$	0.755	0.754
PID4	$\mathbf{Q}_{\mathcal{P}_4}$	$\mathbf{LCC}^T\mathbf{L}^T$	0.800	0.788
PID5	$\mathbf{Q}_{\mathcal{P}_5}$	$\mathbf{LAA}^T\mathbf{L}^T$	0.753	0.752
PID6	$\mathbf{Q}_{\mathcal{P}_6}$	$\mathbf{LXX}^T\mathbf{L}^T$	0.774	0.770
PID7	$\mathbf{Q}_{\mathcal{P}_7}$	$\mathbf{FLL}^T\mathbf{F}^T$	0.777	0.768
PID8	$\mathbf{Q}_{\mathcal{P}_8}$	$\mathbf{FA}^T\mathbf{AF}^T$	0.782	0.778
SID1	$\mathbf{Q}_{\mathcal{S}_1}$	$\mathbf{P}[(\mathbf{CC}^T)\circ(\mathbf{AA}^T)]\mathbf{P}^T$	0.843	0.837
SID2	$\mathbf{Q}_{\mathcal{S}_2}$	$\mathbf{P}[(\mathbf{CC}^T)\circ(\mathbf{XX}^T)]\mathbf{P}^T$	0.832	0.823
SID3	$\mathbf{Q}_{\mathcal{S}_3}$	$\mathbf{L}[(\mathbf{CC}^T)\circ(\mathbf{AA}^T)]\mathbf{L}^T$	0.837	0.829
SID4	$\mathbf{Q}_{\mathcal{S}_4}$	$\mathbf{L}[(\mathbf{CC}^T)\circ(\mathbf{XX}^T)]\mathbf{L}^T$	0.854	0.848
SID5	$\mathbf{Q}_{\mathcal{S}_5}$	$\mathbf{F}[(\mathbf{LL}^T)\circ(\mathbf{A}^T\mathbf{A})]\mathbf{F}^T$	0.820	0.812
ID14	**Combined-kernel (5)**		0.862	0.856
ID15	**Multi-kernel (5)**		**0.883**	**0.875**

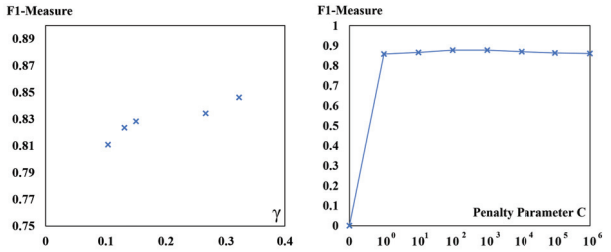

Figure 4: The effectiveness of the proposed method.

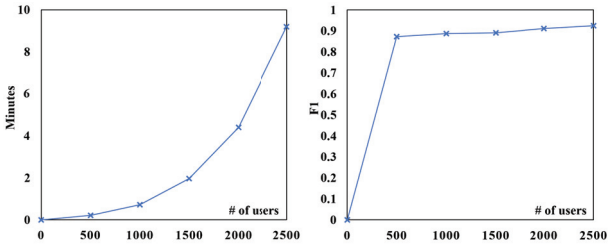

Figure 5: Scalability evaluation of the proposed method.

the weights to form a new kernel (i.e., *ID14*) fed to SVM. From the results shown in Table 2, we can observe that Laplacian score indeed helps us select some important similarities, and the "Combined-kernel (5)" for test set is with 86.2% detection accuracy which works better than any single similarity. This shows that combining different similarities using Laplacian score can improve the performance. However, the method using multi-kernel learning successfully outperforms the single similarities based on different meta-structures and the unsupervised similarity selection algorithm (i.e., Laplacian score), which shows multi-kernel learning can better reflect classification property and thus improve the opioid user detection performance. To demonstrate the effectiveness of multi-kernel learning, we further study the correlation between the learned parameter γ_k weighting each similarity in multi-kernel learner and the actual performance of each similarity in Table 2. From Figure 4: *left*, we can see that the parameter γ_k's can successfully filter out the performance of each similarity. We also further evaluate the parameter sensitivity of combined similarity using multi-kernel learning with different values of the penalty parameter C. From Figure 4: *right*, we can see in a wide range of numbers, the performance of combined similarity is stable and not very sensitive to the penalty parameter. This indicates that for practical use, we can simply tune a parameter using some training data based on cross-validation, and apply that parameter to the test set without concerning the change of the parameter affecting the online performance.

We also further evaluate the training time of our proposed method with different sizes of the training data sets. The scalability is shown in Figure 5: *left*. It is illustrated that the running time is quadratic to the number of training samples. When dealing with more data, approximation or parallel algorithms should be developed. However, as shown in Figure 5: *right*, for such automatic opioid user detection problem, the need of more labels is not as important as the need of more expressive representations of data. The reasons behind this are using HIN representation and meta-structure based approach for relatedness measure over users well describes the rich semantic relationships. Therefore, for practical use, our approach is feasible for real application in automatic opioid user detection.

4.3 Comparisons with other Alternative Methods

In this section, based on the dataset $\mathbf{DB_a}$, we compare our developed system *AutoOPU*, which integrates our proposed method described in Section 3, with three typical classifiers by 10-fold cross-validations, i.e., Naive Bayes (NB), Decision Tree (DT) and SVM. For comparisons, we put content-based information (i.e., user's posted tweet(s) represented by a bag-of-words vector) and all HIN-related relations (i.e., ***R1–R6*** in Section 3.1) as features for different classification methods to learn. The results are shown in Table 3.

From Table 3, we can see that feature engineering helps the performance of machine learning, since the rich semantics encoded in different types of relations can bring more information. However, the use of this information for traditional machine learning algorithms is simply flat features, i.e., concatenation of different features altogether. The results in Table 3 also show that *AutoOPU* further outperforms these alternative classification methods with feature engineering in automatic opioid user detection. To check whether the overall improvement is significant, we also run 20 random trials of training and testing examples to compare *AutoOPU* and SVM with feature engineering, and the probability associated with a paired t-Test (Baldi and Long 2001) with a two-tailed distribution is 4.13×10^{-16}. This shows that *AutoOPU* is significantly better than the best baseline method we compared. The reason behind this is that, in *AutoOPU*, we use more expressive representation for the data,

Table 3: Comparisons with other alternative methods. "Original" means all the methods use original content-based information as features to learn. "Augmented" means that, we combine content-based information and all HIN-related relations as features for different algorithms to learn.

ID	Original	AUC	ACC	F1
1	NB-1	0.678	0.682	0.674
2	DT-1	0.711	0.728	0.727
3	SVM-1	0.724	0.740	0.733
ID	Augmented	AUC	ACC	F1
4	NB-2	0.725	0.731	0.721
5	DT-2	0.766	0.779	0.773
6	SVM-2	0.835	0.837	0.829
7	***AutoOPU***	**0.873**	**0.881**	**0.877**

and build the connection between the higher-level semantics of the data and the final results.

4.4 Case Studies

To better understand opioid addiction epidemic and public perceptions toward MAT, based on the detected opioid users from Twitter using *AutoOPU*, a series of spatio-temporal statistics such as geo-location distribution analysis associated with different timelines and distribution of perceptions and stigmas toward MAT are performed. By making use of the profile data of Tweeter users which indicates their related geo-locations, Figure 6 shows the distribution of the detected opioid users (i.e., 1,132 newly detected heroin users) in different states of the U.S. from Feb. 2016 to Feb. 2017 (the darker color the more severe epidemic the state has).

Similar to the statistics of heroin-related overdose from Centers for Disease Control and Prevention (CDC 2015): Ohio, New York, Illinois, West Virginia and Maryland have larger numbers of heroin users than the others in the U.S. By categorizing all opioid-related tweets posted by the detected opioid users, Figure 6 also shows the distribution of perceptions and stigmas toward MAT in different states of the U.S., where "+" means in favor of MAT (e.g., "*Methadone is an effective treatment which needs to be made more available in the U.S.*") and "-" denotes bias toward MAT (e.g., "*buprenorphine is one drug replaced by another*"). From Figure 6, it can be observed that ***the areas with more severe opioid addiction epidemic also tend to have stigmas toward MAT***. The analysis also indicates that ***there is a remarkable treatment gap suggesting many people who need behavioral health treatment but have not received it due to various reasons*** (e.g., financial and psychological burden). These findings based on the detected opioid users using our developed system *AutoOPU* demonstrate that knowledge from daily-life social media data mining could help sharpen our understanding toward the behavioral process of opioid addiction and treatment.

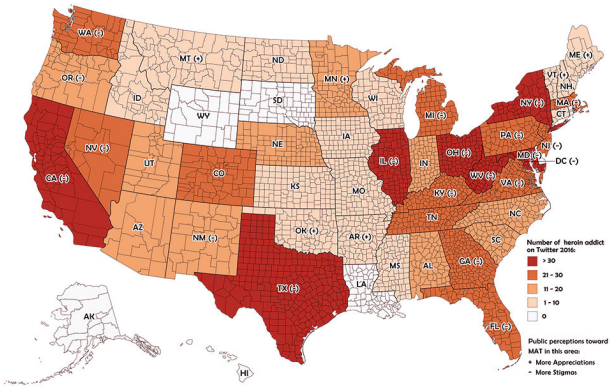

Figure 6: Distribution of heroin users on Twitter in the U.S.

5. Related Work

In recent years, the role of social media in biomedical knowledge mining, such as interactive healthcare and drug pharmacology, has become increasingly important. For example, based on users' posted tweets, a machine learning-based concept extraction system ADRMine was introduced for adverse drug reactions (ADRs) analysis (Nikfarjam et al. 2015); Support Vector Machine (SVM) classifiers based on the content of twitter messages were built to find drug users as well as the potential adverse events (Bian, Topaloglu, and Yu 2012). Unfortunately, the application of social media data analytics into drug-addiction domain has been scarce in the literature with few exceptions (Cameron et al. 2013). Different from the existing works in drug-addition domain, in this paper, we propose to utilize not only the users' posted tweets but also the rich semantic relationships among users, tweets, and topics for opioid user detection from Twitter. Based on the extracted features, the users are represented by a structured HIN.

HIN is used to model different types of entities and relations (Shi et al. 2017). It has been applied to various applications (Sun et al. 2011a; 2011b; Wang et al. 2015; 2016), such as scientific publication network analysis. Several studies have already investigated the use of HIN information for relevance computation, however, most of them only use meta-path (Sun et al. 2011b; Luo et al. 2014) to measure the similarity. Such simple path structure fails to capture a more complex relationship between two entities. To address this problem, Huang et al. (Huang et al. 2016) proposed to use meta-structure, which is a directed acyclic graph of entity and relation types to measure the proximity between two entities. Their work only considered one particular meta-structure to capture the relatedness over entities. Different from existing works, in this paper, we consider different meta-structures which characterize the relatedness over users at different views, and further propose a multi-kernel learning method to aggregate different similarities based on different meta-structures, which is a the first attempt in biomedical knowledge mining.

6. Conclusion

In this paper, we propose a framework called *AutoOPU* to automatically detect opioid users from Twitter. In *AutoOPU*, we first construct a HIN to leverage the information of users, tweets and topics as well as the rich relationships among them. Then, meta-structure based approach is used to characterize the semantic relatedness over users. Afterwards, we integrate content-based similarity and the relatedness depicted by each meta-structure to formulate a similarity measure over users. We then aggregate different similarities using multi-kernel learning for opioid user detection. The promising experimental results on the real data collections from Twitter demonstrate that our framework outperforms other alternative methods.

7. Acknowledgments

This work is supported by WVU NT-NS grant (2016-2017). The work of Y. Zhang, Y. Fan and Y. Ye is partially supported by the U.S. National Science Foundation under grant CNS-1618629.

References

Baldi, P., and Long, A. D. 2001. A bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* 17(6):509–519.

Bian, J.; Topaloglu, U.; and Yu, F. 2012. Towards large-scale twitter mining for drug-related adverse events. In *SHB*, 25–32. ACM.

Blei, D. M.; Ng, A. Y.; Jordan, M. I.; and . 2003. Latent dirichlet allocation. *JMLR* 3(Jan):993–1022.

Cameron, D.; Smith, G. A.; Daniulaityte, R.; Sheth, A. P.; Dave, D.; Chen, L.; Anand, G.; Carlson, R.; Watkins, K. Z.; and Falck, R. 2013. Predose: a semantic web platform for drug abuse epidemiology using social media. *Journal of Biomedical Informatics* 46(6):985–997.

CDC. 2015. *Centers for Disease Control and Prevention, Heroin Overdose Data*. https://www.cdc.gov/drugoverdose/data/heroin.html.

Gönen, M., and Alpaydın, E. 2011. Multiple kernel learning algorithms. *JMLR* 12(Jul):2211–2268.

Han, J.; Sun, Y.; Yan, X.; and Yu, P. S. 2010. Mining knowledge from databases: an information network analysis approach. In *SIGMOD*, 1251–1252. ACM.

He, X.; Cai, D.; and Niyogi, P. 2006. Laplacian score for feature selection. In *NIPS*, 507–514.

Horn, R. A. 1990. The hadamard product. In *Proc. Symp. Appl. Math*, volume 40, 87–169.

Hou, S.; Ye, Y.; Song, Y.; and Abdulhayoglu, M. 2017. Hindroid: An intelligent android malware detection system based on structured heterogeneous information network. In *KDD*, 1507–1515. ACM.

Huang, Z.; Zheng, Y.; Cheng, R.; Sun, Y.; Mamoulis, N.; and Li, X. 2016. Meta structure: Computing relevance in large heterogeneous information networks. In *KDD*, 1595–1604. ACM.

LiveStats. 2013. *TwitterUsageStatistics*. http://goo.gl/hnRg9h.

Luo, C.; Guan, R.; Wang, Z.; and Lin, C. 2014. Hetpathmine: A novel transductive classification algorithm on heterogeneous information networks. In *ECIR*, 210–221. Springer.

Manning, C. D.; Surdeanu, M.; Bauer, J.; Finkel, J. R.; Bethard, S.; and McClosky, D. 2014. The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, 55–60.

Mihalcea, R.; Corley, C.; Strapparava, C.; et al. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI*, volume 6, 775–780.

Murthy, V. H. 2016. Ending the opioid epidemicła call to action. *New England Journal of Medicine* 375(25):2413–2415.

NIDA. 2015. *Overdose Death Rates*. https://goo.gl/tH9Zud.

Nikfarjam, A.; Sarker, A.; OConnor, K.; Ginn, R.; and Gonzalez, G. 2015. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association* 22(3):671–681.

Shi, C.; Li, Y.; Zhang, J.; Sun, Y.; and Philip, S. Y. 2017. A survey of heterogeneous information network analysis. *IEEE TKDE* 29(1):17–37.

Sun, Y., and Han, J. 2012. Mining heterogeneous information networks: principles and methodologies. *Synthesis Lectures on DMKD* 3(2):1–159.

Sun, Z.; Ampornpunt, N.; Varma, M.; and Vishwanathan, S. 2010. Multiple kernel learning and the smo algorithm. In *NIPS*, 2361–2369.

Sun, Y.; Barber, R.; Gupta, M.; Aggarwal, C. C.; and Han, J. 2011a. Co-author relationship prediction in heterogeneous bibliographic networks. In *ASONAM*, 121–128. IEEE.

Sun, Y.; Han, J.; Yan, X.; Yu, P. S.; and Wu, T. 2011b. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *VLDB Endowment* 4(11):992–1003.

Wang, C.; Song, Y.; Li, H.; Zhang, M.; and Han, J. 2015. Knowsim: A document similarity measure on structured heterogeneous information networks. In *ICDM*, 1015–1020. IEEE.

Wang, C.; Song, Y.; Li, H.; and Zhang, J. 2016. Text classification with heterogeneous information network kernels. In *AAAI*, 2130–2136.

Yang, J.; Jiang, Y.-G.; Hauptmann, A. G.; and Ngo, C.-W. 2007. Evaluating bag-of-visual-words representations in scene classification. In *MIR*, 197–206. ACM.