Using Social Media to Combat Opioid Epidemic

Yiming Zhang
ymzhang@mix.wvu.edu

Follow this and additional works at: https://researchrepository.wvu.edu/etd

Part of the Software Engineering Commons

Recommended Citation
Zhang, Yiming, "Using Social Media to Combat Opioid Epidemic" (2019). Graduate Theses, Dissertations, and Problem Reports. 3928.
https://researchrepository.wvu.edu/etd/3928

This Thesis is protected by copyright and/or related rights. It has been brought to you by the The Research Repository @ WVU with permission from the rights-holder(s). You are free to use this Thesis in any way that is permitted by the copyright and related rights legislation that applies to your use. For other uses you must obtain permission from the rights-holder(s) directly, unless additional rights are indicated by a Creative Commons license in the record and/or on the work itself. This Thesis has been accepted for inclusion in WVU Graduate Theses, Dissertations, and Problem Reports collection by an authorized administrator of The Research Repository @ WVU. For more information, please contact researchrepository@mail.wvu.edu.
USING SOCIAL MEDIA TO COMBAT OPIOID EPIDEMIC

Yiming Zhang

Thesis submitted to the
Benjamin M. Statler College of Engineering and Mineral Resources
at West Virginia University
in partial fulfillment of the requirements
for the degree of
Master of
Science in
Computer Science

Yanfang Ye, Ph.D.
Xin Li, Ph.D.
Elaine M. Eschen, Ph.D.

Lane department of computer science and electrical engineering
Morgantown, West Virginia
2018

Keywords: Social Media; Opioid User Detection;
Heterogeneous Information Network

Copyright 2018 Yiming Zhang
ABSTRACT

Using Social Media to Combat Opioid Epidemic

Yiming Zhang

Opioid addiction has become one of the largest and deadliest epidemics in the United States. To combat such deadly epidemic, there is an urgent need for novel tools and methodologies to gain new insights into the behavioral processes of opioid abuse and addiction. The role of social media in biomedical knowledge mining has turned into increasingly significant in recent years. The data from social media may contribute information beyond the knowledge of domain professionals (e.g., psychiatrists and epidemiology researchers) and could potentially assist in sharpening our understanding toward the behavioral process of opioid addiction and treatment.

In this thesis, we propose a novel framework to automate the analysis of social media (i.e., Twitter) for the detection of the opioid users. To model the Twitter users and posted tweets as well as their rich relationships, we constructed a structured heterogeneous information network (HIN) for representation. We then introduce a meta-path-based approach to characterize the semantic relatedness over users. As different meta-paths depict the relatedness over users at different views, we used Laplacian scores to aggregate different similarities formulated by different meta-paths and then a transductive classification model was built to make predictions. We conduct a comprehensive experimental study based on the real sample collections from Twitter to validate the effectiveness of our proposed approach. To improve the performance of automatic opioid user detection, we presented a meta-structure-based method to depict relatedness and integrate content-based similarity to formulate a similarity measure over users. We then aggregate different similarities using multi-kernel learning for opioid user detection. Comprehensive experimental results on real sample collections from Twitter demonstrate the effectiveness of our proposed learning models.
To my families
Acknowledgments

I would first like to express my greatest gratitude to my committee chair and advisor, Dr. Yanfang Ye, for her guidance and support not only for this thesis but throughout the time of my whole master study. Her passion, vision, attitude, and love for research is always an inspiration source and influence to me; her expertise, understanding, generous guidance, suggestions, valuable comments and revisions make it possible for me to work on such an exciting topic; her devotion of significant time and efforts on mentoring my research has resulted in seven publications by the date of this thesis.

I would also like to thank my committee members, Dr. Li and Dr. Eschen, for their time and help for my research work; I am very fortunate to work with a cheerful group members, including Yujie Fan, Shifu Hou, Lingwei Chen, Jian Liu and Aaron Saas, who exchanged ideas about machine learning related research work and provided useful suggestions on my thesis.

I am highly thankful to be blessed by amazing and talented family members and friends, who have made such a positive impact on my daily life, study, and research.
# Table of Contents

Abstract .................................................. ii
Acknowledgments ......................................... iv
List of Figures ............................................ vi
List of Tables ............................................. vii

**Chapter 1 Introduction** .................................. 1
   1.1 Background and Motivation .......................... 1
   1.2 Research Objective ................................ 2
   1.3 Major Contributions ............................... 3
   1.4 Organization of the Thesis ....................... 4

**Chapter 2 Related Work** ................................ 5

**Chapter 3 Proposed Method** ............................ 7
   3.1 Heterogeneous Information Network Construction .... 7
   3.2 Meta-path and Meta-structure Based Similarities ...... 9
   3.3 Transductive Classification Built on Meta-path Based Similarities ... 12
   3.4 Multi-kernel Learning Built on Meta-structure Based Similarities ... 14

**Chapter 4 System Architecture** ......................... 17

**Chapter 5 Experimental Results And Analysis** ........... 19
   5.1 Data Collection and Annotation .................... 20
   5.2 Evaluation of Meta-path and Meta-structure Based Similarities .... 20
   5.3 Comparisons with Traditional Machine Learning Methods .... 22
   5.4 Scalability and Stability Evaluations ................ 23
   5.5 Case Studies ...................................... 24

**Chapter 6 Conclusion and Future Work** ................... 27

Publications .............................................. 28
Bibliography ............................................. 28
# List of Figures

3.1 Network schema. ................................................. 8  
3.2 Meta-paths and meta-structures. ................................. 9  
4.1 System architecture of our proposed framework. ................. 17  
5.1 Scalability evaluation. ............................................ 23  
5.2 Stability evaluation. .............................................. 23  
5.3 Distribution of heroin users on Twitter. ......................... 24  
5.4 Identification of the influential users ............................ 26
List of Tables

3.1 The description of each meta-path. ........................................ 10
3.2 The description of each meta-structure. ................................ 11

5.1 Performance indices of opioid user detection. ....................... 19
5.2 Evaluation of meta-path and meta-structure based similarities. ...... 21
5.3 Comparisons with traditional machine learning methods .......... 22
5.4 Deep understanding of detected heroin users. ....................... 25
Chapter 1

Introduction

1.1 Background and Motivation

Opioids are a group of drugs which include the illegal drug heroin and powerful pain relievers by legal prescription, such as morphine and oxycodone. Opioid addiction has become one of the largest and deadliest epidemics in the United States [1]. Americans are more likely to die of a drug overdose than in a motor vehicle accident and overdose deaths have increased every subsequent year [2]. In 2016, 11.8 million Americans age 12 and up were reported current non-medical use of prescription opioids [3]. There was a skyrocketing increase of opioid related death in the past decade: according to National Institute on Drug Abuse (NIDA), in 2017, 49,068 Americans died involving opioid overdose and 15,958 people died from heroin overdose, both reflecting significant increase from 2001 [4]. 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 [5]. 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 [6]. 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 [7] and interactive healthcare [8], has become increasingly impor-
tant. Due to the growing use of the Internet, never-ending growth of data are generated from the social media offering 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 [9]. A large-scale Twitter users are willing to share their experiences of using opioids (e.g., “I have a crippling heroin addiction and it’s 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’s 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.

1.2 Research Objective

To achieve the goal, in this thesis, we propose a novel framework named AutoDOU to automate the detection of opioid users from Twitter, where meta-path [10] based on heterogeneous information network (HIN) are used to characterize the relatedness over users. As the moral says “man is known by the company he keeps”, to detect if a user is an opioid user, we not only analyze the posted tweets but also his/her social network. For example, a user posted a tweet “I’ll bring some heroin”, which might not be sufficient to determine if he/she is an opioid user. However, with the information that one of his/her tweeps (i.e., Twitter people that follow each other) @ him/her in the tweet “Let’s shoot heroin tonight hahahah. where’s the needles at?”, we can conclude that the user is highly possible an opioid user. To model the users and posted tweets as well as their rich semantic relationships, a structured heterogeneous information network (HIN) [11], which is capable to be composed of different types of entities and relations, is first introduced. Then we use meta-path [10] to incorporate higher-level semantics to build up the relatedness of users. In this way, a similarity between two users can not only capture whether they are posting similar tweets but also capture whether they have strong social relations, such as post tweets discuss the same topic, post tweets mention the same user. Since there can be multiple meta-paths to define different similarities, we incorporate all useful meta-paths with their weights computed by Laplacian scores [12].
To reduce the cost of acquiring labeled samples for supervised learning, we construct a transductive classification model [13] to detect the opioid users based on HIN and the combined meta-path.

Although meta-path has been shown to be useful for relatedness measure between users [10, 14], 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. To improve the performance of automatic opioid user detection, in this thesis, we propose another 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. In AutoOPU, 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 [15] based approach to characterize the semantic relatedness over users. Then, we further integrate content-based similarity (i.e., similarity of users’ posted tweets) and relatedness depicted by each meta-structure to formulate a similarity measure over users. Later, we aggregate different similarities using multi-kernel learning [16], each of which is automatically weighted by the learning algorithm to make predictions.

1.3 Major Contributions

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 frameworks are 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 the meta-path/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-path/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 *transductive classification model in HIN* for opioid user detection. Inductive classification has been applied in HIN to predict the unlabeled entities. However, it usually requires large number of labeled data to achieve better accuracy. In other words, when training data decreases, its detection accuracy may greatly compromise. In our application, obtaining the labeled data (either opioid users or non-opioid users) from Twitter is both time-consuming and cost-expensive. To overcome this challenge, we present a transductive classification model in HIN to reduce the cost of acquiring labeled samples for opioid user detection.

- 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 *two practical systems* AutoDOU and 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 systems in comparisons with traditional machine learning approaches.

### 1.4 Organization of the Thesis

The remainder of this paper is organized as follows. Chapter 2 discusses the related work. Chapter 3 presents our proposed method in detail. Chapter 4 introduces our system architecture. In Chapter 5, based on the real sample collections and annotations from Twitter, we systematically evaluate the performance of our methods. Finally, Chapter 6 concludes.
Chapter 2

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 [17]; 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 [18]. Unfortunately, the application of social media data analytics into drug-addiction domain has been scarce in the literature with few exceptions: Cameron et al. [19] developed a novel semantic web platform called PREDOSE (Prescription Drug Abuse Online Surveillance and Epidemiology) to facilitate the epidemiologic study of prescription and related drug abuse practices using social media; Sarker et al. [20] designed an automatic supervised classification technique to distinguish posts containing signals of medication abuse. However, most of these studies merely used content-based features (e.g., posted tweets or messages) for their applications. Actually, the relations among users and the generated contents are also very important for target user detection. Different from the existing works in drug-addition domain, in this paper, we propose to not only utilize users’ posted tweets but also the relationships among users and tweets (i.e., user-user, user-tweet, tweet-tweet, tweet-topic relations) for opioid user detection from Twitter. Based on the extracted features, the users are represented by a structured heterogeneous information network (HIN), meta-path and meta-structure based approaches are used to link the users.
Heterogeneous information network (HIN) has been intensively studied in recent years. Typically, HIN is used to model different types of entities and relations [11]. It has been applied to various applications, such as scientific publication network analysis [21, 10] and document analysis based on knowledge graph [22]. Different from traditional graph similarities, such as shortest path, the similarity defined on HIN, i.e., PathSim [10], is more likely a natural extension to dot product. Different from the simple dot product, the similarity defined over HIN considers the semantics of the network metadata. In our work, to measure the similarities over users, we develop a similarity based on multiple meta-paths using an unsupervised meta-path weighting mechanism [22].

To solve the classification problem in HIN, compared with inductive methods [23, 24], transductive classification [25, 26, 27, 14] was proposed to reduce the cost of acquiring labeled samples in supervised learning. However, it has yet applied in biomedical knowledge mining. In this paper, we explore how to construct an effective transductive classification model in HIN for opioid user detection from Twitter. To address the problem that simple path structure (i.e., meta-path) fails to capture a more complex relationship between two entities., Huang et al. [15] proposed to use meta-structure, which is a directed acyclic graph of entity and relation types to measure the proximity between two entities. However, their work only considered one particular meta-structure to capture the relatedness over entities. Different from their works, 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.
Chapter 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 Heterogeneous Information Network Construction

As the above discussion, to detect opioid users from Twitter, we not only utilized users’ posted tweets but also the rich semantic 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 $P$ where each element $p_{i,j} \in \{0, 1\}$ denotes if user $i$ posts tweet $j$.

- **R2**: To denote the relation that a user likes a tweet, we generate the user-like-tweet matrix $L$ 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 $F$ 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 $A$ 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 $X$ 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 $C$ where each element $c_{i,j} \in \{0,1\}$ indicates if tweet $i$ contains topic $j$. In our application, we use Latent Dirichlet allocation [28] 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) [29].** A HIN is defined as a graph $G = (V,E)$ with an entity type mapping $\phi: V \rightarrow A$ and a relation type mapping $\psi: E \rightarrow R$, where $V$ denotes the entity set and $E$ is the relation set, $A$ denotes the entity type set and $R$ is the relation type set, and the number of entity types $|A| > 1$ or the number of relation types $|R| > 1$.

![Network schema](Link_to_image)

**Figure 3.1: Network schema.**

**Definition 2. Network schema [10].** The network schema for a HIN $G$, denoted as $\mathcal{T}_G = (A, R)$, is a graph with nodes as entity types from $A$ and edges as relation types from $R$.  

8
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 3.1.

### 3.2 Meta-path and Meta-structure Based Similarities

The different types of entities and different relations between them motivate us to use a machine-readable representation to enrich the semantics of similarities among users. Meta-path [10] is used in the concept of HIN to formulate the semantics of higher-order relationships among entities. Here we follow this concept and extend it for the detection of opioid users.

**Definition 3. Meta-path [10].** A meta-path $\mathcal{P}$ is a path defined on the graph of network schema $\mathcal{T}_G = (A, R)$, 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 $\cdot$ denotes relation composition operator, and $L$ is the length of $\mathcal{P}$.

![Figure 3.2: Meta-paths and meta-structures.](image)

In our case, based on the HIN schema displayed in Figure 3.1, we generate eight meaningful meta-paths to characterize the relatedness over users (i.e., $PID1$–$PID8$ shown...
For example, \textit{PID1} depicts that two users are related if they have posted tweets discussed same topics; while \textit{PID2} denotes that two users are related by their posted tweets mentioning same users. To compute entity similarities using a particular meta-path, we use the following commuting matrix [10] to give a general form.

**Definition 4. Commuting matrix** [10]. Given a network $G = (V, E)$ and its network schema $T_G$, a commuting matrix $M_P$ for a meta-path $P = (A_1A_2...A_{L+1})$ is defined as $M_P = G_{A_1A_2} G_{A_3A_4} ... G_{A_{L-1}A_L}$, where $G_{A_iA_j}$ is the adjacency matrix between types $A_i$ and $A_j$. $M_P(i,j)$ represents the number of path instances between entity $x_i \in A_1$ and entity $y_i \in A_{L+1}$ under meta-path $P$.

For the former example, the adjacently matrix between users and tweets is $G_{user,tweet}$. Then the commuting matrix of users computed using the meta-path $\text{user } \text{post} \rightarrow \text{tweet} \text{ contain } \text{topic}$ is $\text{PCC}^T \text{PT}$, whose element denotes the number of tweet pairs posted by this pair of users that discuss the same topics. Given a network schema with different types of entities and relations, we can enumerate a lot of meta-paths. In our works, based on the collected data, resting on the six different kinds of relationships, we construct eight meaningful meta-paths as listed in Table 3.1 for similarity measures over users.

<table>
<thead>
<tr>
<th>PID</th>
<th>Matrix $M$</th>
<th>Description of each element $m_{ij}$ in $M$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\text{PCC}^T \text{PT}$</td>
<td># of tweet pairs posted by user $i$ and $j$ that contain same topics</td>
</tr>
<tr>
<td>2</td>
<td>$\text{PAA}^T \text{PT}$</td>
<td># of tweet pairs posted by user $i$ and $j$ that mention same people</td>
</tr>
<tr>
<td>3</td>
<td>$\text{PXX}^T \text{PT}$</td>
<td># of tweet pairs posted by user $i$ and $j$ that repost same tweets</td>
</tr>
<tr>
<td>4</td>
<td>$\text{LCC}^T \text{LT}$</td>
<td># of tweet pairs liked by user $i$ and $j$ that contain same topics</td>
</tr>
<tr>
<td>5</td>
<td>$\text{LAA}^T \text{LT}$</td>
<td># of tweet pairs liked by user $i$ and $j$ that mention same people</td>
</tr>
<tr>
<td>6</td>
<td>$\text{LXX}^T \text{LT}$</td>
<td># of tweet pairs liked by user $i$ and $j$ that repost same tweets</td>
</tr>
<tr>
<td>7</td>
<td>$\text{FLL}^T \text{FT}$</td>
<td># of tweet pairs of user $i$ and $j$ who like same tweets</td>
</tr>
<tr>
<td>8</td>
<td>$\text{FA}^T \text{AT}$</td>
<td># of tweet pairs of user $i$ and $j$ who are mentioned in same tweets</td>
</tr>
</tbody>
</table>

Although meta-path has been shown to be useful for relatedness measure between users [10, 14], 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 [15] 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 [15]:

**Definition 5. Meta-structure** [15]. A meta-structure $S$ is a directed acyclic graph with a single source node $n_s$ and a single target node $n_t$, defined on a HIN schema $T_G = (A, R)$. Formally, $S = (N, M, n_s, n_t)$, where $N$ is a set of nodes and $M$ is a set of edges. For any node $x \in N$, $x \in A$; for any link $(x, y) \in M$, $(x, y) \in R$.

<table>
<thead>
<tr>
<th>SID</th>
<th>Commuting matrix $\mathbf{M}$</th>
<th>Description of each element $m_{ij}$ in $\mathbf{M}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\mathbf{P}[(\mathbf{CC}^T) \circ (\mathbf{AA}^T)]\mathbf{P}^T$</td>
<td># of tweet pairs posted by user $i$ and $j$ that contain same topics and mention same people</td>
</tr>
<tr>
<td>2</td>
<td>$\mathbf{P}[(\mathbf{CC}^T) \circ (\mathbf{XX}^T)]\mathbf{P}^T$</td>
<td># of tweet pairs posted by user $i$ and $j$ that contain same topics and repost same tweets</td>
</tr>
<tr>
<td>3</td>
<td>$\mathbf{L}[(\mathbf{CC}^T) \circ (\mathbf{AA}^T)]\mathbf{L}^T$</td>
<td># of tweet pairs liked by user $i$ and $j$ that contain same topics and mention same people</td>
</tr>
<tr>
<td>4</td>
<td>$\mathbf{L}[(\mathbf{CC}^T) \circ (\mathbf{XX}^T)]\mathbf{L}^T$</td>
<td># of tweet pairs liked by user $i$ and $j$ that contain same topics and repost same tweets</td>
</tr>
<tr>
<td>5</td>
<td>$\mathbf{F}[(\mathbf{LL}^T) \circ (\mathbf{A}^TA)]\mathbf{F}^T$</td>
<td># of tweet pairs of user $i$ and $j$ who like same tweets and are mentioned in same tweets</td>
</tr>
</tbody>
</table>

Based on the HIN schema displayed in Figure 3.1, we generate five meaningful meta-structures to characterize the relatedness over users (i.e., $\text{SID}_1$–$\text{SID}_5$ shown in Figure 3.2: right). For example, $\text{SID}_1$ 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 $\text{SID}_4$ 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., $\text{PID}_1$ and $\text{PID}_2$ are particular cases of $\text{SID}_1$). In Figure 3.2, the meta-paths of $\text{PID}_1$–$\text{PID}_8$ (left) are the special cases of the constructed meta-structures $\text{SID}_1$–$\text{SID}_5$ (right). But meta-structure is capable to express more complex relationship in a convenient way.

To measure the relatedness over users using a particular meta-structure designed above, we use commuting matrix to compute the counting-based similarity matrix for a meta-structure. Take $\text{SID}_1$ as an example, the commuting matrix of users computed using $\text{SID}_1$ is $\mathbf{P}[(\mathbf{CC}^T) \circ (\mathbf{AA}^T)]\mathbf{P}^T$, where $\mathbf{P}$, $\mathbf{C}$, $\mathbf{A}$ are the adjacency matrices between two
corresponding entity types, \( \odot \) denotes the Hadamard product of two matrices. Whose element \( M_{S_1}(i,j) \) denotes the number of tweet pairs posted by user \( i \) and user \( j \) which contain same topics and also mention same people. Table 3.2 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-path or meta-structure to form a similarity measure matrix over users. The similarity matrix over users is denoted as \( Q \), whose element is the combination of content-based similarity and relation based similarity. We define similarity matrix \( Q \) based on \( M_{P_k} \) or \( M_{S_k} \) as:

\[
Q_{P_k}(i,j) = [1 + \log(M_{P_k}(i,j) + 1)] \times tSim(i,j),
\]

\[
Q_{S_k}(i,j) = [1 + \log(M_{S_k}(i,j) + 1)] \times tSim(i,j),
\]

where \( M_{P_k}(i,j) \) is the relatedness between user \( i \) and \( j \) under meta-path \( P_k \), \( M_{S_k}(i,j) \) is the relatedness between user \( i \) and \( j \) under meta-structure \( S_k \), \( tSim(i,j) \) 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 [30] to estimate the closeness of two users’ posted content.

### 3.3 Transductive Classification Built on Meta-path Based Similarities

Different meta-paths measure the similarities between two users at different views. Instead of using a single meta-path for similarity measure over two users, we propose to combine different meta-paths and weight each of them for user classification (i.e., whether he/she is an opioid user). Suppose there are \( K \) meta-paths \( P_k \) with their corresponding commuting matrices \( M_{P_k} \), \( k = 1, 2, ..., K \), we use Eq.(3.1) to compute the similarity matrix \( Q_{P_k} (k = 1, 2, ..., K) \) based on \( M_{P_k} \). Following [22], after the normalization of each similarity matrix, we combine different meta-paths to form a new
similarity measure:

\[
Q'(i, j) = \frac{2 \times \sum_{k=1}^{K} w_k Q_{pk}(i, j)}{\sum_{k=1}^{K} w_k Q_{pk}(i, i) + \sum_{k=1}^{K} w_k Q_{pk}(j, j)},
\]

(3.3)

where \( w = [w_1, w_2, ..., w_K] \) is the weighted vector of the meta-paths. In our works, we use Laplacian score [12] to learn the weight of each meta-path, since it can be computed to reflect the locality preserving power of each feature.

Compared with inductive classification methods [23, 24] which only use objects with known labels for training, transductive classification models [13, 31] can also utilize the relatedness between objects to propagate labels and thus reduce the cost of acquiring labeled data for training. In recent years, transductive classification algorithms have been devised in HIN [26] for the applications such as classifying the bibliographic data into research communities [25]. In our case, since it is time-consuming and cost-expensive to obtain the labeled data (either opioid users or non-opioid users) from Twitter, we propose to use transductive classification in HIN for opioid user detection. We first introduce the concept of transductive classification in HIN as follow.

Definition 6. Transductive classification in HIN [25]. Given an HIN \( G = (V, E) \) with \( m \) types of entities \( V_i \subseteq \bigcup_{i=1}^{m} V_i \), where \( V_i \subseteq A_i \) (\( i = 1, ..., m \)). Suppose \( V' \) is a subset of \( V \) which is with class labels of \( C = \{C_1, ..., C_c\} \), where \( c \) is the number of classes. The classification task is to predict the labels for all the unlabeled entities \( V - V' \).

In transductive classification model, there are two assumptions of consistency: Assumption (1) – entities with tight relationship tend to have a high possibility being in the same class; and Assumption (2) – the classification results should consist with the prelabeled information. Following these two assumptions, learning with local and global consistency algorithm (LLGC) in homogeneous information network was proposed in [13] for classification. In our application, we further extend the LLGC framework to classify the entities in heterogeneous information network for opioid user detection, whose cost function can be denoted as follow:

\[
Q(F) = \frac{1}{2} \left( \sum_{i,j=0}^{n} Q'(i, j) \| \frac{F_i}{\sqrt{D_{ii}}} - \frac{F_j}{\sqrt{D_{jj}}} \|^2 + \mu \sum_{i=1}^{n} \| F_i - Y_i \|^2 \right),
\]

(3.4)

where \( n \) is the number of entities (i.e., Twitter users) in HIN, \( M' \) is the similarity matrix
combing different meta-paths, $F$ is a $n \times c$ ($c$ is the number of classes) matrix whose element $F(i, j)$ represents the possibility of user $i$ belonging to class $j$, $Y$ is also a $n \times c$ matrix containing the prelabeled information, $D$ is a diagonal matrix whose $(i, i)$-element is equal to the sum of the $i$-th row of $Q'$, and $\mu > 0$ is the regularization parameter. The first term of the right-hand side in Eq.(3.4) is called smoothness constraint satisfying Assumption (1), which means a good classifying function should not change too much between nearby points; the second term is fitting constraint following Assumption (2), which indicates a good classifying function should not change too much from the initial label assignment. The parameter $\mu$ captures the trade-off between these two competing constraints. Note that the fitting constraint contains both labeled and unlabeled data. In our application, to initialize the matrix $Y$ [32], a classier is trained (i.e., SVM) resting on the content-based features (i.e., tweets posted by labeled users) to assign an initial label for each unlabeled entity (i.e., user) in HIN.

Based on Eq.(3.4), the classifying function can be defined as

$$F^* = \arg \min Q(F). \quad (3.5)$$

To obtain the optimal $F$, we differentiate $Q(F)$ with respect to $F$ and then have

$$\frac{\partial Q}{\partial F} = F^* - SF^* + \mu(F^* - Y), \quad (3.6)$$

where $S = D^{-1/2}M'D^{-1/2}$. Eq.(3.6) can be further transformed into [13]

$$F^* = \beta(I - \alpha S)^{-1}Y, \quad (3.7)$$

where $\alpha = \frac{1}{1 + \mu}$, $\beta = \frac{\mu}{1 + \mu}$.

### 3.4 Multi-kernel Learning Built on Meta-structure Based Similarities

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 \( S_k \) \((k = 1, 2, ..., K)\), we can calculate their corresponding commuting matrices \( M_{S_k} \) \((k = 1, 2, ..., K)\). Then, we use Eq.(3.2) to compute the similarity matrix \( Q_{S_k} \) \((k = 1, 2, ..., K)\) based on \( M_{S_k} \). We treat each similarity matrix \( Q_{S_k} \) as a kernel in multi-kernel learning model. If the matrix \( Q_{S_k} \) 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 [16, 33]:

\[
Q' = \sum_{k=1}^{K} \gamma_k Q_{S_k},
\]

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 \( \{x_i, y_i\}_{i=1}^{N} \), where \( x_i \) 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 [16] with following objective function for parameter learning:

\[
\min_{w>0, \xi_i, \gamma_k \geq 0} \frac{1}{2} \sum_{k} \|w_k\|^2/\gamma_k + C \sum_i \xi_i + \frac{\lambda}{2} \left( \sum_k \gamma_k^p \right)^{2/p},
\]

\[
s.t. \quad y_i \left( \sum_k w_k^T \varphi_k(x_i) + b \right) \geq 1 - \xi_i,
\]

where \( w_k \) is a weight vector associated with each kernel. For each data \( \{x_i, y_i\} \), the slack parameter \( \xi_i \) is introduced to allow mis-classification. \( \varphi_k(x_i) \) is a nonlinear mapping function of features in the Hilbert space that defines the kernel, where \( Q_{S_k}(i,j) = \varphi_k(x_i)^T \varphi_k(x_j) \). Then by applying the representation theorem, we have \( w_k = \sum_i \alpha_i \varphi_k(x_i) \). \( \alpha_i \) can be solved using the dual formulation, and non-zero \( \alpha_i \)'s lead to the support vectors. For another set of parameters \( \gamma_k \), the \( p \)-norm \( (\sum_k \gamma_k^p)\) 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 \( \gamma_k \)'s are obtained to reveal the
importance of different similarities based on different meta-structures. For a user \( x \),

\[
\sum_k w_k \varphi_k(x) + b,
\]

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

**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 \( S_S = \{ S_1, S_2, ..., S_5 \} \) based on the six matrices above; 
3: for each meta-structure \( S_k \) \((k = 1, 2, ..., 5)\) in \( S_S \) do 
4: Compute \( M_{S_k} \) based on Definition 5; 
5: Compute \( Q_{S_k} \) using Eq.(3.1); 
6: end for 
7: Let each \( Q_{S_k} \) be a kernel in the multi-kernel learning model, and compute the weight vector \( w_k \) for each kernel by optimizing Eq.(3.9); 
8: for each user \( x \) in \( T_e \) do 
9: Predict its label using Eq.(3.10); 
10: end for
Chapter 4

System Architecture

Figure 4.1 shows the system architecture of our proposed framework including two systems AutoDOA and AutoOPU for automatic opioid user detection from Twitter, which consists of the following major components.

1. **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 [34].

2. **Feature Extractor and HIN Constructor.** A bag-of-words [35] feature vector will be extracted to represent each user’s posted tweet(s). Besides, the relation-
ships among users, tweets and topics will be further analyzed, such as, i) user-follow-user, ii) user-like-tweet, iii) tweet-mention-user, iv) tweet-RT-tweet, and v) tweet-contain-topic. Based on the extracted features, a structural HIN is then constructed to represent users. (See Section 3.1 for details.)

3. **System-1: AutoDOA.** The AutoDOA consists of two major sub-components, meta-path builder and transductive classifier.

   - **Meta-path Builder.** In this module, different meta-paths associated with their corresponding commuting matrices are generated from HIN to measure the similarities between users. Laplacian scores are further computed to weight the importance of different meta-paths. Given the weighted meta-paths, the different commuting matrices is combined to formulate a more powerful similarity measure over users. (See Section 3.2 for details.)

   - **Transductive Classifier.** To reduce the cost of acquiring labeled samples for supervised learning, a transductive classification model in HIN is built to automatically detect the opioid users. (See Section 3.3 for details.)

4. **System-2: AutoOPU.** Different from AutoDOA, the AutoOPU is composed of meta-structure builder and multi-kernel Learner.

   - **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 3.4 for details.)

5. **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.

18
Chapter 5

Experimental Results And Analysis

In this section, we show three sets of experimental studies using real sample collections from Twitter to fully evaluate the performance of our developed system AutoDOA and AutoOPU for automatic detection of opioid users: (1) in the first set of experiments, we evaluate the effectiveness of meta-path and meta-structure based similarities; (2) in the second set of experiments, we evaluate the proposed methods by comparisons with traditional classification methods; and (3) in the third set of experiments, we evaluate the scalability and stability of our developed systems for opioid user detection. Table 5.1 shows the measures for evaluation of different methods.

Table 5.1: Performance indices of opioid user detection.

<table>
<thead>
<tr>
<th>Indices</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP</td>
<td># correctly classified as opioid users</td>
</tr>
<tr>
<td>TN</td>
<td># correctly classified as non-opioid users</td>
</tr>
<tr>
<td>FP</td>
<td># mistakenly classified as opioid users</td>
</tr>
<tr>
<td>FN</td>
<td># mistakenly classified as non-opioid users</td>
</tr>
<tr>
<td>Precision</td>
<td>$\frac{TP}{TP + FP}$</td>
</tr>
<tr>
<td>Recall</td>
<td>$\frac{TP}{TP + FN}$</td>
</tr>
<tr>
<td>ACC</td>
<td>$\frac{(TP + TN)}{(TP + TN + FP + FN)}$</td>
</tr>
<tr>
<td>F1</td>
<td>$2 \times \frac{Precision \times Recall}{Precision + Recall}$</td>
</tr>
</tbody>
</table>
5.1 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 DBa) 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).

5.2 Evaluation of Meta-path and Meta-structure Based Similarities

In this set of experiments, based on the annotated dataset DBa and the HIN schema (described in Section 3.1), we fully evaluate the effectiveness of meta-path and meta-structure based similarities in opioid user detection. In the set of experiments, we randomly select 90% of the data for training, while the remaining 10% is used for testing.

We first construct eight meta-paths (i.e., PID1–PID8 shown in Figure 3.2: left) and five meta-structures (i.e., SID1–SID5 shown in Figure 3.2: right). To measure the similarities over users, as described in Section 3.2, we integrate similarities of users’ posted tweets and relatedness depicted by each meta-path or meta-structure to form a similarity measure matrix. We evaluate their performances for opioid user detection using Support Vector Machine (SVM). For each meta-path or meta-structure, 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 10.
Table 5.2: Evaluation of meta-path and meta-structure based similarities.

<table>
<thead>
<tr>
<th>ID</th>
<th>Kernel</th>
<th>Commuting Matrix</th>
<th>ACC</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>PID1</td>
<td>$Q_{P1}$</td>
<td>$PCC^TP^T$</td>
<td>0.806</td>
<td>0.792</td>
</tr>
<tr>
<td>PID2</td>
<td>$Q_{P2}$</td>
<td>$PAA^TP^T$</td>
<td>0.773</td>
<td>0.768</td>
</tr>
<tr>
<td>PID3</td>
<td>$Q_{P3}$</td>
<td>$PXX^TP^T$</td>
<td>0.755</td>
<td>0.754</td>
</tr>
<tr>
<td>PID4</td>
<td>$Q_{P4}$</td>
<td>$LCC^TL^T$</td>
<td>0.800</td>
<td>0.788</td>
</tr>
<tr>
<td>PID5</td>
<td>$Q_{P5}$</td>
<td>$LAA^TL^T$</td>
<td>0.753</td>
<td>0.752</td>
</tr>
<tr>
<td>PID6</td>
<td>$Q_{P6}$</td>
<td>$LXX^TL^T$</td>
<td>0.774</td>
<td>0.770</td>
</tr>
<tr>
<td>PID7</td>
<td>$Q_{P7}$</td>
<td>$FLL^TF^T$</td>
<td>0.777</td>
<td>0.768</td>
</tr>
<tr>
<td>PID8</td>
<td>$Q_{P8}$</td>
<td>$FA^TAF^T$</td>
<td>0.782</td>
<td>0.778</td>
</tr>
<tr>
<td>ID9</td>
<td>Combined-kernel (8)</td>
<td>/</td>
<td>0.836</td>
<td>0.827</td>
</tr>
<tr>
<td>SID1</td>
<td>$Q_{S1}$</td>
<td>$P[(CC^T) \circ (AA^T)]P^T$</td>
<td>0.843</td>
<td>0.837</td>
</tr>
<tr>
<td>SID2</td>
<td>$Q_{S2}$</td>
<td>$P[(CC^T) \circ (XX^T)]P^T$</td>
<td>0.832</td>
<td>0.823</td>
</tr>
<tr>
<td>SID3</td>
<td>$Q_{S3}$</td>
<td>$L[(CC^T) \circ (AA^T)]L^T$</td>
<td>0.837</td>
<td>0.829</td>
</tr>
<tr>
<td>SID4</td>
<td>$Q_{S4}$</td>
<td>$L[(CC^T) \circ (XX^T)]L^T$</td>
<td>0.854</td>
<td>0.848</td>
</tr>
<tr>
<td>SID5</td>
<td>$Q_{S5}$</td>
<td>$F[(LL^T) \circ (A^T A)]F^T$</td>
<td>0.820</td>
<td>0.812</td>
</tr>
<tr>
<td>ID15</td>
<td>Combined-kernel (5)</td>
<td>/</td>
<td>0.862</td>
<td>0.856</td>
</tr>
</tbody>
</table>

The results in Table 5.2 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 also evaluate the combined similarity [22] of all the constructed meta-paths (i.e., PID1–PID8) and meta-structures (i.e., SID1–SID5) using Laplacian scores as their weights [12] to form two new kernel (i.e., ID9 and ID14) fed to SVM. From the results shown in Table 5.2, we can observe that Laplacian score indeed helps us select some important similarities, and the “Combined-kernel (8)” and “Combined-kernel (5)” for test set are with 83.6% and 86.2% detection accuracy which works better than their related single similarity. This shows that combining different similarities depicted by different meta-paths or meta-structures using Laplacian score can further improve the
performance, since it not only utilizes content-based features but also diverse relation-based features which include rich semantic information in opioid user detection.

## 5.3 Comparisons with Traditional Machine Learning Methods

In this section, based on the dataset $DB_a$, we randomly select a portion of the labeled data (range from 90% to 50%) to simulate the experiments. We compare our developed systems $AutoDOA$ and $AutoOPU$ with three typical classifiers i.e., Naive Bayes (NB), Decision Tree (DT) and Support Vector Machine (SVM). For comparisons, we combine 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.

<table>
<thead>
<tr>
<th></th>
<th>90%</th>
<th>80%</th>
<th>70%</th>
<th>60%</th>
<th>50%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NB</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACC</td>
<td>0.7245</td>
<td>0.6963</td>
<td>0.6687</td>
<td>0.6448</td>
<td>0.6115</td>
</tr>
<tr>
<td>F1</td>
<td>0.7103</td>
<td>0.6884</td>
<td>0.6636</td>
<td>0.6319</td>
<td>0.6068</td>
</tr>
<tr>
<td><strong>DT</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACC</td>
<td>0.7569</td>
<td>0.7245</td>
<td>0.7060</td>
<td>0.6674</td>
<td>0.6387</td>
</tr>
<tr>
<td>F1</td>
<td>0.7206</td>
<td>0.7013</td>
<td>0.6740</td>
<td>0.6478</td>
<td>0.6118</td>
</tr>
<tr>
<td><strong>SVM</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACC</td>
<td>0.8336</td>
<td>0.8167</td>
<td>0.7752</td>
<td>0.7426</td>
<td>0.7021</td>
</tr>
<tr>
<td>F1</td>
<td>0.8215</td>
<td>0.8002</td>
<td>0.7536</td>
<td>0.7241</td>
<td>0.6956</td>
</tr>
<tr>
<td><strong>AutoDOA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACC</td>
<td>0.8578</td>
<td>0.8454</td>
<td>0.8297</td>
<td>0.8087</td>
<td>0.8016</td>
</tr>
<tr>
<td>F1</td>
<td>0.8448</td>
<td>0.8369</td>
<td>0.8242</td>
<td>0.8015</td>
<td>0.7967</td>
</tr>
<tr>
<td><strong>AutoOPU</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACC</td>
<td>0.8816</td>
<td>0.8586</td>
<td>0.8299</td>
<td>0.7971</td>
<td>0.7687</td>
</tr>
<tr>
<td>F1</td>
<td>0.8776</td>
<td>0.8513</td>
<td>0.8240</td>
<td>0.7943</td>
<td>0.7619</td>
</tr>
</tbody>
</table>

The experimental results are illustrated in Table 5.3. From Table 5.3, we can see when training data decreases (from 90% to 50%), $AutoDOA$ using transductive classification model over HIN and combined meta-path based similarities works better than other methods in automatic opioid user detection, since the detection performances of $AutoDOA$ (based on transductive classification model) don’t change too much (i.e., both ACC and F1 drop less than 5%); while the detection performances of $AutoOPU$, NB, DT and SVM (based on inductive classification model) were greatly compromised as...
training samples decrease (i.e., both ACC and F1 drop more than 10%). This is because that AutoDOA not only uses the information from training data for prediction but also utilizes the relatedness among training samples and testing objects to propagate labels. However, when there are a larger proportion of training data (i.e., 90% or 80%), AutoOPU significantly outperforms other methods including AutoDOA and three baseline methods in automatic opioid user detection. The reason behind this is that, in AutoOPU, we use multi-kernel leaning model built on HIN and combine different meta-structure based similarities which have more expressive representation for the data, and build the connection between the higher-level semantics of the data and the final results.

5.4 Scalability and Stability Evaluations

Based on the dataset DB_a, we systematically evaluate the performance of our developed systems AutoDOA and AutoOPU, including the detection scalability and stability.

We first evaluate the training time of AutoDOA and AutoOPU with different sizes of the training data sets. Figure 5.1 shows the scalability of our proposed methods. 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 5.2, 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 reason behind this is our methods using HIN representation, meta-path and meta-structure based approaches for relatedness measure over users well describes the rich semantic relationships. Therefore, for practical use, our approaches are feasible for real application in automatic opioid user detection.

Figure 5.1: Scalability evaluation.  
Figure 5.2: Stability evaluation.
5.5 Case Studies

In this section, after the automatic detection of opioid users from Twitter using our proposed framework, to better understand opioid addiction epidemic and public perceptions toward Medication-Assisted Treatment (MAT), we further analyze the data and conduct some case studies based on the detected opioid users.

Figure 5.3: Distribution of heroin users on Twitter.

- **Case study 1: epidemic surveillance of opioid abuse and addiction in the U.S.** To better understand the distribution and opioid addiction epidemic, a series of spatio-temporal statistics such as geo-location distribution analysis associated with different timelines are performed based on our detected opioid users. By making use of the profile data of Tweeter users which indicates their related geo-locations, Figure 5.3 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 [36]: Ohio, New York, Illinois, West Virginia and Maryland have larger numbers of heroin users than the others in the U.S. Though in some rural areas Twitter is not the primary platform for social communication and may have
its biases, this case study still clearly reflects the actual status of opioid addiction epidemic in the U.S., which demonstrates that using social media for epidemic surveillance of opioid abuse and addiction is practical and feasible.

- **Case study 2: deep understanding of public perceptions toward MAT.** To further assess the public perceptions and stigmas of MAT, we randomly sample 30% of our detected opioid users (i.e. 356 heroin users) to further analyze and categorize their posted tweets. Table 5.4 shows different categories of the posted tweets as well as two cases of public perceptions toward MAT. The study reveals that (1) some users show appreciations of MAT (e.g., “Methadone is an effective treatment which needs to be made more available in the U.S.”), while (2) some of them still have significant stigmas toward MAT who mistakenly think of MAT as “one drug replaced by another”. It also shows that there is a remarkable treatment gap suggesting the majority of people who need behavioral health treatment but have not received it due to various reasons (e.g., public stigma, financial burden).

Table 5.4: Deep understanding of detected heroin users.

<table>
<thead>
<tr>
<th>Categories of the posted tweets</th>
<th># tweets</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Need heroin</td>
<td>130</td>
<td>36.52%</td>
</tr>
<tr>
<td>Shoot heroin</td>
<td>103</td>
<td>28.92%</td>
</tr>
<tr>
<td>Love heroin</td>
<td>82</td>
<td>23.03%</td>
</tr>
<tr>
<td>Bought heroin</td>
<td>13</td>
<td>3.65%</td>
</tr>
<tr>
<td>Reasons of heroin addiction</td>
<td>11</td>
<td>3.09%</td>
</tr>
<tr>
<td>Attitude toward heroin addiction</td>
<td>8</td>
<td>2.25%</td>
</tr>
<tr>
<td>Perceptions toward MATs</td>
<td>5</td>
<td>1.40%</td>
</tr>
<tr>
<td>Consequences of heroin addiction</td>
<td>3</td>
<td>0.84%</td>
</tr>
<tr>
<td>Seek for help</td>
<td>1</td>
<td>0.28%</td>
</tr>
</tbody>
</table>

Examples of perceptions toward MAT

1. **Appreciation for MAT:** “heroin; I think this model of treatment (methadone) needs to be made available in the US, as it’s the most effective treatment for opioid.”
2. **Stigma toward MAT:** “heroin mat is utterly fraudulent but expensive. You can treat all you want, the addicts will go right back to it.”

- **Case study 3: identification of the influential users to advertise the best practice of MAT.** To promote the perception of MAT, we believe it is best to first
locate the group of users with apparent stigmas toward MAT and then use social network analysis to identify the most likely authoritative users that could influence the group of interests. In this study, the assumption is further validated. For the users who post their perceptions of MAT, we further analyze their social networks (e.g., their tweeps and people who like/repost/reply their tweets) and find that they actively interact with their virtual friends on Twitter, which indicates that they could be the influential users who have the power of authoritative sources in the linked environment and thus can help promote the perception of MAT. Figure 5.4 shows two examples of potential influential users who can help advertise the best practice of MAT.

Figure 5.4: Identification of the influential users

The above case studies based on the automatically detected opioid users using our developed systems demonstrate that knowledge from daily-life social media data mining could support a better practice of opioid addiction prevention and treatment.
Chapter 6

Conclusion and Future Work

In this paper, we propose two frameworks called AutoDOA and AutoOPU to automatically detect opioid users from Twitter. In AutoDOA, we first construct a heterogeneous information network (HIN) to leverage the information of users and tweets as well as the rich relationships among them, which gives the user a higher-level semantic representation. Then, Laplacian scores are computed to weight different generated meta-paths and a combined meta-path is used for similarity measure over users. To reduce the cost of acquiring labeled samples, a transductive classification model in HIN is then built for opioid users detection. In AutoOPU, the 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 frameworks outperform other alternate methods. The case studies also show that knowledge from daily-life social media data mining could support a better practice of opioid addiction prevention and treatment.

In our future work, we will continue to improve our system to automate analysis in other social media (e.g., Facebook, Instagram, Reddit, etc) for biomedical knowledge mining. On the other hand, the study such as computational cost and incremental learning over heterogeneous information networks is still worth exploring.
Publications

1. Yujie Fan, **Yiming Zhang**, Yanfang Ye, Xin Li. “Automatic Opioid User Detection from Twitter: Transductive Ensemble Built on Different Meta-graph Based Similarities over Heterogeneous Information Network.” *In IJCAI, 2018.* (20.5% acceptance rate)


Bibliography


