Harnessing Social Media to Identify Homeless Youth At-Risk of Substance Use

Zi-Yi Dou,[1] **Anamika Barman-Adhikari,**[2] **Fei Fang,**[1] **Amulya Yadav**[3]
[1]Carnegie Mellon University
[2]University of Denver
[3]Pennsylvania State University
zdou@cs.cmu.edu, Anamika.BarmanAdhikari@du.edu, feif@cs.cmu.edu, amulya@psu.edu

Abstract

Homeless youth are a highly vulnerable population and report highly elevated rates of substance use. Prior work on mitigating substance use among homeless youth has primarily relied on survey data to get information about substance use among homeless youth, which can then be used to inform the design of targeted intervention programs. However, such survey data is often onerous to collect, is limited by its reliance on self-reports and retrospective recall, and quickly becomes dated. The advent of social media has provided us with an important data source for understanding the health behaviors of homeless youth. In this paper, we target this specific population and demonstrate how to detect substance use based on texts from social media. We collect 135K Facebook posts and comments together with survey responses from a group of homeless youth and use this data to build novel substance use detection systems with machine learning and natural language processing techniques. Experimental results show that our proposed methods achieve ROC-AUC scores of 0.77 on identifying certain kinds of substance use among homeless youth using Facebook conversations only, and ROC-AUC scores of 0.83 when combined with answers to four survey questions that are not about their demographic characteristics or substance use. Furthermore, we investigate connections between the characteristics of people's Facebook posts and substance use and provide insights about the problem.

Introduction

An estimated 1.5 to 3 million youth aged 18-25 experience homelessness each year in the United States (Toro, Lesperance, and Braciszewski 2011), reporting elevated rates of hard drug use, including street opioids (such as heroin), prescription opioids, and stimulants (such as cocaine, crack, and methamphetamines) (Nyamathi et al. 2012; Kennedy et al. 2010). For example, studies have reported lifetime rates of use ranging from 24 percent to 69 percent for methamphetamines (meth) and 12 percent to 15 percent for heroin (Nyamathi et al. 2012) among homeless youth.

Social media sites such as Facebook present an important data source for understanding the social context of homeless youths' health behaviors, including substance use behaviors. Preliminary studies with these youth find surprisingly pervasive social media usage: ∼80 percent of homeless youth report using social media weekly (only 9 percent do not have a social media profile), and ∼90 percent of youth who used social media prefer to use Facebook (Barman-Adhikari et al. 2016; Guadagno, Muscanell, and Pollio 2013). More importantly, studies suggest that youth's online interactions may be associated with their substance use behaviors (Rice, Milburn, and Monro 2011). Barman-Adhikari et al. (2016) find that almost one-third of youth report talking about drug consumption on social media.As these youth are transient and difficult to engage in place-based services (i.e., a physical location), social media may represent a novel venue for screening and intervening upon homeless youth to prevent and reduce their substance use behaviors.

In fact, research already suggests that social media may be even more important for youth who are homeless because it can be a significant resource for a population that is traditionally lacking in resources (Jones and Fox 2009). For example, research has found that youth who are homeless use social media for instrumental purposes: 28 percent use it to locate housing and 13 percent use it to look for social services (Rice et al. 2013). Thus, there is potential to leverage social media to understand substance use among homeless youth and provide interventions to those in need of help. To achieve this goal, it is critical to first identify homeless youth at-risk of substance use, a task that is typically done by collecting lengthy survey data from the target population, which is time-consuming, has limited reliance due to self-reports and retrospective recall, and it quickly becomes dated.

To the best of our knowledge, no published work exists that has collected observational social media data on this extremely high-risk group of young adults to predict their patterns of substance use. In this paper, we address this limitation by collecting ∼135K social media posts and comments that a group of homeless youth post within one year, together with rich survey responses that indicate their substance use behavior. We apply machine learning algorithms to identify the ones who are at-risk of substance use by (i) using their social media content only; and (ii) combining their social media content with their answers to only four survey questions that are unrelated to their demographic characteristics or substance use. To collect social media posts, we rely on Facebook as that is by far the most popular social media platform among homeless youth. Hopefully, our developed methods can lead to tools that are easy-to-use and can 1) help

identify homeless youth who may need intervention even if they are not willing to answer lengthy surveys truthfully; and 2) can potentially be integrated into an online service tool so that it can reach and help more homeless youth.

From a machine learning perspective, this problem is challenging for a variety of factors. First, different from traditional text classification problems, Facebook posts are typically noisy, as they can contain a significant amount of typographical errors or Internet slangs. Besides, as is often the case, only a limited amount of data is available and thus we need to develop robust algorithms that have robust performance in order to deal with these scarce-data settings. As a result of these challenges, several novel adaptations are required if we want to apply established machine learning algorithms to predict substance use for homeless youth.

To address these challenges, we develop a general-purpose multi-step framework which consists of multiple steps of pre-processing and vectorization, followed by a combination of late-fusion and early-fusion techniques for effective training of predictive models on scarce and noisy social-media data. Experimental results demonstrate the effectiveness of the proposed methods, achieving ROC-AUC scores of $\sim$0.77 on identifying certain kinds of substance use among homeless youth with social media conversations only, and scores of $\sim$0.8 when combined with answers to four survey questions. In addition, we investigate associations between certain characteristics of people's Facebook posts and substance use and provide several unique insights about the problem.

In short, we make the following novel contributions:

- We target a highly vulnerable population, i.e., homeless youth, which has received little attention in previous work, and collect Facebook data and survey responses from them (a first-of-its-kind effort to the best of our knowledge).
- We investigate several substance use detection models with machine learning and natural language processing techniques that are specifically adapted to noisy social media texts, and exhibit superior performance of our models on real-world data by solely using Facebook posts, and further improvements can be achieved when non-drug-related survey responses are included as additional inputs.
- We investigate associations between certain characteristics of word usages (or survey responses) and substance use, and gain insights about the problem.
- We demonstrate that our proposed methods can benefit the homeless youth by presenting specific use cases regarding substance use prevention in real world settings.

Related Work

Social Media and Health-Risk Behaviors among Homeless Youth. Only three studies (Rice, Monro, and Barman-Adhikari 2010; Rice, Milburn, and Monro 2011; Barman-Adhikari et al. 2016) have assessed social media use and health-risk behaviors among homeless youth. In (Rice, Monro, and Barman-Adhikari 2010), $\sim$25% of the homeless youth surveyed report looking for a sex partner online and are more likely to engage in exchange sex or survival sex. However, none of them analyze how to leverage social media data to detect the group of homeless youth at risk of substance use and our work fills in the gap.

Mining Information from Social Media Texts. With the advent of social media, users have a tendency to post a large quantity of data online, including what they have done and how they feel, which has been used to study users' behaviors. For instance, Aramaki, Maskawa, and Morita (2011) extract information from Twitter to detect influenza epidemics and Gerber (2014) uses linguistic analysis and statistical topic modeling to automatically identify discussion topics for predicting different types of crime. Researchers have tried to understand behaviors of substance users or predict substance use using social media data. Zhou et al. (2016) understand behaviors of illicit drug users by collecting Instagram posts and utilizing a dictionary of illicit drug-related slangs to find common substance use behaviors with regard to time. Ding, Bickel, and Pan (2017) explore several ways to predict whether a user suffers from substance use disorder, and we treat it as one of our baselines.

While machine learning has been used to predict substance use among other youth and adult populations, researchers have not applied them to a group of young adults who face transient living circumstances and also experience very high rates of trauma which translate to extremely high rates of substance use engagement. Besides, all the previous methods fail to consider the noise in social media texts, which have been shown to degrade the model performance. In addition, we ask participants to complete a survey on their demographic information and health conditions, and propose methods that outperform previous state-of-the-art algorithms by utilizing both Facebook posts and survey responses to detect substance use among homeless youth.

Dataset

We collected a total of 135,189 textual Facebook conversations (posts and comments) from 158 survey participants (homeless youth) who shared this content on their Facebook profiles during the data collection period. A purposive sampling design (a non-probabilistic sampling method which uses a pre-defined list of characteristics for the population based on the objective of the study) was used to recruit participants. Recruiters were present at a non-profit agency, over six months, for the duration of service provision hours to approach and screen youth, and invite participation.

Youth who were interested in the study were screened for eligibility. Eligibility criteria was assessed by a trained research assistant who asked participants about: where they slept last night, how long could they stay at that location, their age, and whether they owned a Facebook profile for at least a year. For youth who met eligibility criteria, the research assistant sought informed consent for participation. For those who did not meet eligibility, the research assistant thanked them for their time and discontinued the interaction. For eligible participants, we collected all Facebook posts shared by them in the last one year. The resulting dataset consists of $\sim$135K posts in total.

In the effort of pre-processing the data for our analysis, we removed the Facebook posts and comments that are either empty or only contain weblinks from our dataset. The resulting dataset consists of 91,482 Facebook conversations, including 24,960 Facebook posts and 66,522 comments.

In addition to collecting their Facebook information, we also asked participants to fill out a self-reported survey that collected information such as their demographic information, past and current living status, etc. In addition to people's basic demographic information such as age and gender, the participants were also asked questions like "Why did you leave home or become homeless?" and "How often do you feel that you lack companionship?" that are not directly about demographic characteristics or substance use, yet they can be utilized for substance use predictions. Table 1 summarizes the general aspects of the survey data and its participants' Facebook conversations revealed from their posts and comments. Because not all the participants have shared both their Facebook conversations and filled out the survey, we removed users who either do not have Facebook posts available or do not complete the survey, resulting in a dataset consisting of $\sim$25K posts and $\sim$66K comments from 87 Facebook profiles.

Most importantly, in the survey, the participants were asked to note if they have used drugs in the last 30 days. Specifically, they reported whether they have used marijuana, cocaine (including powder, coke, blow or snow), crack (including freebase or rock), heroin, methamphetamines, ecstasy, needles to inject any illegal drug into their body, and/or prescription drugs without a doctor's prescription or more often than prescribed. The statistics of people using drugs are shown in Table 2. We hope a machine learning model has the ability to predict which specific drug one person is using based on Facebook posts, survey responses, or both. Prior research suggests that there are unique predictors and consequences of some kind of drugs compared to others. For example, a recent study found that drug users who used methamphetamine had an 80 percent greater risk of attempting suicide than drug users who did not (Marshall et al. 2011). Also, homeless youth who use illicit drugs experience longer episodes of homelessness and victimization while living on the streets compared to recreational drugs such as alcohol, tobacco, and marijuana (Bender et al. 2015). Because we can see from the table that only two people were using crack, we only ran experiments on the other seven types of drugs. Note that while collecting this survey data is onerous in day-to-day settings, it is important from an evaluation perspective, as the survey allows us to gather ground truth labels for our prediction task, i.e., which people are substance users and which people are not. In addition, we also analyze the impact of these survey questions on our predictive performance, and discuss potential workarounds in the paper below.

Methods

In this section, we will discuss our algorithms in detail. We first describe each component of our algorithms separately, including the pre-processing and vectorization steps, and then illustrate the overall classification procedure.

Preprocessing Social Media Texts

Noise in social media text is a known issue that has been investigated in a variety of previous work (Michel and Neubig 2018), with most of them focusing on data augmentation. Unfortunately, we empirically show that popular data augmentation methods do not work sufficiently well in our problem domain. As a result, in this paper, we approach the problem from a completely different perspective, and propose a general-purpose methodology that utilizes subword information for handling noise in social media text.

Subwords are an effective solution to the out-of-vocabulary problem, which is commonly observed in noisy social media text. In this work, we employ byte pair encoding (BPE) to perform subword segmentation, which is a simple data compression technique that is widely used in machine translation (Sennrich, Haddow, and Birch 2016). The basic idea of BPE is to iteratively merge pairs of characters or character sequences that appear frequently in the corpus to create subwords.

In order to compare against data-augmentation based methods for handling noise in social media data, we also attempted to utilize BPE-Dropout, a recently proposed data augmentation technique, to tackle the problem. The procedure of BPE-Dropout is simple and it mainly alters the segmentation procedure of BPE while keeping its original merge table. At a high level, BPE-Dropout stochastically corrupts the segmentation procedure of BPE, which can benefit the machine learning models by 1) augmenting the dataset; 2) enabling them to be robust against noise. Because both of these properties can be of great benefit in our setting, BPE-Dropout seems to be a promising technique to use.

Vector Representations for Users

After pre-processing the inputs (using BPE), we also need to vectorize the inputs before feeding them to machine learning (ML) models, such that these vector representations can be processed by ML models. This process is referred to as vectorization. In this part, we describe how we obtain vectorized representations for each homeless youth (or user), ranging from a simple bag-of-words model to a more complicated distributed bag-of-words model.

Bag-of-Words Model The bag-of-words model (BoW) is a simple and intuitive vectorization method for text classification. The idea of the bag-of-words model is to convert text into fixed-length vectors by counting how many times each word (or subword) appears in the input text. One caveat of the bag-of-words model is that it does not take word order into consideration. However, we empirically demonstrate that the method is effective (see experiments).

Singular Value Decomposition Typically, a vocabulary can contain thousands of entries, which can cause bag-of-words (BoW) models to break down without sufficient amounts of training data. To solve this problem, we use Singular Value Decomposition (SVD) to reduce the dimensions of the BoW vector representation. SVD (De Lathauwer, De Moor, and Vandewalle 2000) is a matrix factorization method that generalizes the eigendecomposition of a square

Data Sources	Characteristics	Mean	Standard Deviation	Min	Max
Survey (158 observations)	age	20.72	1.89	18	25
	#posts	243.44	275.96	1	1598
	#comments	133.61	189.64	1	1061
Posts (21,179 observations)	#characters	108.53	174.07	4	1452
Comments (12,025 observations)	#characters	59.71	92.93	4	1452

Table 1: Summary of survey and survey participants' Facebook conversation data.

	Marijuana	Cocaine	Crack	Heroin	Metha.	Ecstasy	Needles	Prescription
#positive cases	61	10	2	4	18	7	7	10
positive cases per drug type (%)	70.11	11.49	2.30	4.60	20.69	8.05	8.05	11.49

Table 2: Number and percentage of positive cases.

matrix to any $m \times n$ matrix. Concretely, given any $m \times n$ matrix A, the SVD algorithm can find matrices U, W and V which satisfy the equation $A = UWV^T$, where U is an orthogonal matrix with a size of $m \times n$, W is a $n \times n$ diagonal matrix and V is an $n \times n$ orthogonal matrix. To perform dimension reduction after obtaining matrices U, V, W by SVD, we first keep the r ($r < n$) largest singular values in the diagonal matrix W and obtain the resulting matrix W', then compute a new matrix $A' = UW'$.

In our setting, the variable m is the number of homeless youth in the data. We treat all the posts from each user as one big document and utilize the BoW model to vectorize each user's post. The vectors of all the users are then concatenated to form the matrix A, which is reduced to a new matrix A' of size $m \times r$ using SVD. Each row of the new matrix A' becomes the new feature vector for each user. The new features will have significantly fewer dimensions than the old ones and words with similar meanings can share similar representations. However, valuable information might be lost during compression and thus we need to find a balance between efficiency and effectiveness.

Document Embedding with Distributed Bag-of-Words Model Previously, researchers have tried to learn document embeddings with distributed memory and the distributed bag-of-words model (D-DBoW) (Le and Mikolov 2014). The idea of these approaches is simple: during training, either a document vector and one or more word vectors are aggregated to predict a target word in the context, or a document vector is fed to a neural network to predict words randomly sampled from the document.

Specifically, we treat all the posts by one user as one document as before, and try to train a document vector to represent each user. At each training step, a global document vector v_i will be sampled, which is treated as the representation of the i-th user. Then, we sample n (sub)words from the posts uploaded by the i-th user. A neural network is trained to maximize the likelihood of the n sampled words given the global document representation v_i. The training process will be repeated until convergence. After the training completes, we can get vector representations for all the users. Following Ding, Bickel, and Pan (2017), we choose the document embedding with distributed bag-of-words model (D-DBoW) approach in our drug use prediction domain.

Multi-Task Learning

When the amount of training data is limited, multi-task learning can be used to add additional supervision to the model, as well as function as a regularizer. In our setting, we have information on which types of drugs are being used by each homeless youth in our dataset. Therefore, when we predict whether a homeless youth (user) is consuming one type of drug (say type A), we can utilize information about alternate types of drugs that they may have consumed (in addition to drug type A) and make predictions simultaneously.

While additional supervision signals from related tasks can be typically helpful, multi-task learning with unrelated task objectives can be harmful and deteriorate the model performance. To alleviate this issue, instead of predicting all types of drugs simultaneously, for each type of drug, we try to perform multi-task learning with different combinations of types of drugs and evaluate the model performance on a development set. The combination which achieves the best performance will be selected to train the model. While this brute-force algorithm can be computationally inefficient, it can ensure that we only utilize the positive connections.

Multi-View Learning

As we have illustrated in the previous section, in our settings, we not only have access to the Facebook posts from all the homeless youth, but also have them complete a questionnaire that documents their biographic information as well as their answers to multi-choice questions like "How would you rate your perceived health?". In order to utilize information from the questionnaire and combine it with the posts, we propose both early fusion and late fusion techniques.

Early Fusion. The idea of the early fusion strategy is to concatenate the features of posts and the features of the questionnaires into a single vector before feeding them to classifiers. After the concatenation, the classifier is trained using techniques as before.

Index	Methods	Marijuana	Cocaine	Heroin	Meth.	Ecstasy	Inject	Prescription	Average
1	Majority Voting	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
Using Posts									
2	Ding, Bickel, and Pan (2017)	0.503	0.435	0.476	0.530	0.503	0.532	0.519	0.500
3	BERT (Devlin et al. 2019)	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
4	BOW	0.532	0.435	0.491	0.523	0.533	0.468	0.545	0.504
5	SVD	0.450	0.538	0.436	0.452	0.532	0.532	0.479	0.488
6	D-DBoW	0.583	0.485	0.464	0.592	0.515	0.552	0.454	0.521
7	D-DBoW + Multi	0.617	**0.655**	0.796	0.711	**0.774**	0.622	0.712	0.698
Using Survey Answers									
8	Survey	0.440	0.632	0.491	0.636	0.500	0.464	0.626	0.541
Combining Posts and Comments									
9	Early Fusion	0.668	0.594	0.724	0.641	0.706	0.713	0.648	0.671
10	Late Fusion	0.681	0.617	0.732	0.632	0.723	0.779	**0.747**	0.702
Combining Posts, Survey Answers and Comments									
11	Early Fusion	**0.702**	**0.680**	0.750	0.633	0.694	0.728	**0.747**	0.704
12	Late Fusion	0.677	0.648	**0.815**	**0.712**	0.694	**0.826**	0.728	**0.728**

Table 3: ROC-AUC socres of models. The highest scores are in bold.

Late Fusion. The late fusion strategy first trains separate classifiers on different views, then ensembles the classifiers together. Concretely, different classifiers will be trained on different views. When making the final predictions, all the classifiers' outputs will be combined. In this paper, we have attempted a meta-classifier approach. The meta-classifier approach takes the output probability of both classifiers as input (one classifier for each view) and outputs the final probability, and is trained with the training samples.

Overall Algorithm

First, we use subword segmentation algorithms to segment words from Facebook posts to subwords. Afterwards, we convert the inputs into vectors by applying the bag-of-words, SVD, and document embedding with distributed bag-of-words algorithms. The converted vectors are then fed into a machine learning classifier, which is trained with multi-task learning. We also use one-hot representations for vectorizing survey answers. Then, for each user, we either use the early fusion or the late fusion algorithms to do classifications. We choose decision tree as the base classifier in this paper, and we also try other options (see experiments).

Experiments

We evaluate our models on our collected dataset. In this section, we first describe our basic experimental settings and baselines we compare our models with, then present the experimental results. We also conduct a fair amount of ablation studies and analysis to gain some insights to our model.

Settings

Datasets. The detailed description of our collected dataset can be found in the Dataset section. We removed users who do not have either survey responses or Facebook posts, resulting in 87 datapoints containing information from $\sim$25K Facebook posts and $\sim$66k comments. We lowercased all the texts after performing subword segmentation strategies, which resulted in 8K merge operations. For BPE-Dropout, we performed the algorithm three times with a dropout rate of 0.1, resulting in a dataset that is three times larger than the original dataset. It should be noted that we did not perform any kind of tokenizations for subword-level models. We performed an analysis on a validation set and chose survey answers to the multi-choice questions "Why did you leave home or become homeless?", "How often do you feel that you lack companionship?", "I can share happy and sad moments with these friends", "In your first 18 years of life, a parent or other adult in the household often pushed, grabbed, slapped or threw something at you" because the answers to these questions are correlated with substance use and they are not directly about demographic characteristics or substance use.

Implementation Details. Because of the scarcity of the data, we mainly tried decision tree instead of deep neural networks for classification. We have also attempted random forests to perform classifications as they have also shown to be powerful models on small-scale datasets and can be easily adapted to multi-learning classification settings. The feature size is set to 50 For SVD and D-DBoW.

Baselines. We compared our model with three baselines: 1) majority voting; 2) the strongest model in Ding, Bickel, and Pan (2017); 3) BERT (Devlin et al. 2019), which achieves state-of-the-art performance on a variety of tasks. In addition, we conducted extensive ablation studies to

	Marijuana	Cocaine	Heroin	Methamphetamines	Ecstasy	Needles	Prescription	Average
D-DBoW	0.583	0.485	0.464	0.592	0.515	0.552	0.454	0.521
-subword	0.503	0.435	0.476	0.530	0.503	0.532	0.519	0.500
+BPE-Dropout	0.513	0.464	0.509	0.601	0.482	0.485	0.519	0.512
+random forest	0.519	0.487	0.500	0.500	0.500	0.589	0.553	0.521
+#post	0.480	0.429	0.491	0.457	0.532	0.474	0.552	0.488

Table 4: Ablation studies on D-DBoW.

demonstrate the necessity of each component in our algorithmic framework. We used three-fold cross validation and weighted ROC-AUC scores to evaluate the performance of the model.

Main Results

Single-View Learning We first trained models with data from only one view. As we can see from row 1-7 in Table 3, our models are consistently better than all the baseline models. BERT cannot outperform the simple majority voting strategy, probably because BERT is mainly trained with Wikipedia data, and the huge domain differences between Wikipedia articles and social media can cause the degraded performance of the model. The strongest method in Ding, Bickel, and Pan (2017) (row 2) can outperform the majority voting mechanism to some extent, while being outperformed by our models. As the main difference between Ding, Bickel, and Pan (2017) and our methods is the adoption of the subword model, the improvements indicate the necessity of using subword models in our settings. Despite its simplicity, the bag-of-words model (row 4) can achieve reasonable performance compared with the simple baseline models. The SVD model (row 5), however, cannot improve upon the baseline in most cases, possibly because the compression can drop some valuable information that can be useful for classification. The D-DBoW method (row 6), on the other hand, can improve the baseline by a large margin, which is consistent with the previous findings (Ding, Bickel, and Pan 2017). Multi-task learning (row 7) is highly beneficial in our setting, as it outperforms all the other methods significantly, which demonstrates the effectiveness of additional supervision from other tasks. Specifically, in the best scenarios, the model can achieve a ROC-AUC score of 0.774, improving the next best baseline by 0.241 points. Using survey answers alone (row 8) can achieve over 0.6 ROC-AUC scores on three types of drugs, which is quite effective compared to other methods. The results are intuitive, as information such as people's emotional stability and social life can reveal if they are engaging in substance use.

Multi-View Learning As we have described above, a natural idea is to combine information from different views. In our settings, we have access to people's posts and comments on Facebook, as well as their survey answers, and we have tried to combine information from these "views". As we can see from the results (row 9-10) in Table 3, combining posts and comments does not always help, probably because there is some overlap between comments and posts. As demonstrated in row 11-12, because there is greater diversity among survey answers and social media texts, adding survey answers to the fusion can improve the model performance, with the best AUC-ROC score being 0.826. In addition, comparing both early fusion and late fusion strategies, we can find that late fusion generally outperforms early fusion, indicating that combining high-level information is better than fusing low-level features in our settings.

Ablation Studies

We also did a fair amount of ablation studies and the results are shown in Table 4. First, we take the D-DBoW model and try not to use subword segmentation. Instead, we use the Twitter-aware tokenizer in the NLTK package[1] (which is designed to be flexible and easy to adapt to new domains and tasks) to segment posts into sequences of words. We can see from the table that this modification significantly degrades the performance, which shows that the use of subword segmentation algorithms is necessary. As mentioned in the the method section, we also attempted to utilize the BPE-Dropout algorithm. As shown in the table, surprisingly, the adoption of the BPE-Dropout algorithm would lead to degraded performance. We conjecture that this is because the augmented datapoints are similar to the original ones, which can cause the model to overfit the training data. Next, we tried to use random forests as our classification algorithm (instead of decision trees). However, we can see that the adoption of random forests does not improve the model performance. One possible explanation is that since the dimension of features is small (< 30), and predictions of every tree in the random forest are correlated with each other, a combination of these trees can result in a relatively poor generalization ability. We also tried to provide the number of posts as one feature to the model. However, this leads to degraded performance, which suggests that this may not be a reasonable feature in our setting.

Analysis

Associations between Word Usages/Survey Answers and Substance Use We first check the associations between word usages or survey answers and substance use. It should be noted that here we just classify people into two types, namely substance-users and non-substance users, without considering which specific type of substance they are using. We compute the correlations by directly training a linear SVM classifier with the number of appearances of one

[1] https://www.nltk.org/api/nltk.tokenize.html

	Non-Substance User	Substance User
Words	sincerely (0.611), love (0.549), ...	sucking (0.535), '.' (0.526)
Survey Answers	I can share my happy and sad moments with friends (0.611)	In your first 18 years of life, a parent or other adult in the household often push, grab, slap or throw something at you, or ever hit you so hard you had marks or were injured. (0.611)
Sentences	"So cute!! I want one!! ¡3." I miss my Daughter My favorite person in the world.	"They either don't know, don't understand or don't care." Smoke weed every day

Table 5: Associations between words (or survey answers) and substance use.

	anger	anticipation	disgust	fear	joy	sadness	surprise	trust
Substance Users	0.419	0.505	0.325	0.346	0.544	0.331	0.252	0.594
Non-Substance Users	0.353	0.473	0.300	0.325	0.593	0.312	0.219	0.582

Table 6: Sentiments of posts from substance users and non-substance users.

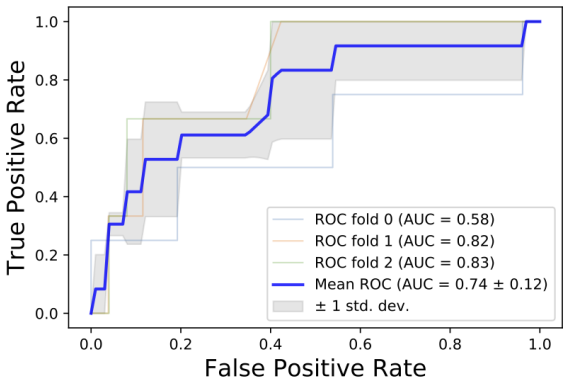

Figure 1: The ROC curve of our algorithm.

specific word (or survey answer) as input features, and compute the resulting ROC-AUC scores. If the weight of the SVM classifier is positive, then the correlation is positive, and otherwise the correlation is negative. The correlations as well as the correlation scores are shown in Table 5. We can see that positive words like "sincerely", "love" can partially indicate that a person is not a substance user. Similarly, negative words might be an indicator that a person is likely to engage in substance use. Survey questions that can convey people's current emotional and social status, such as "I can share my happy and sad moments with friends", is highly correlated with non-substance use. From the answers to this question, we can tell if a person has close friends and if they are able to express their emotions in a reasonable way, which might be an indicator of substance use.

Sentiments We use a lexicon-based sentiment analysis tool[2] to analyze the sentiments of each post. From Table 6, we can see that posts from substance users can exhibit more negative feelings such as anger, disgust, sadness, whereas non-substance users post more positive things and their posts contain more joyful sentiments.

Facebook Post Examples We display some post examples that can be associated with the use of drugs in Table 5. Sometimes people's posts can directly convey if they are using substances. For example, some people may post "smoke weed every day" on Facebook. In some other cases, texts that can indicate good relationships with their friends, family can be associated with positive sentiments and indicate that a person is not a substance user, and vice versa.

Balancing Between True and False Negative Rate We train a model that predicts whether a user is using substances or not and plot the ROC curve as shown in Figure 1. We can see that by tuning the prediction threshold, we can actually balance between true and false negative rates. In applications where we may have limited resources, we may want a high true positive rate and we can set the threshold to a high value.

Conclusion

We collect Facebook conversations and survey responses from homeless youth and develop machine learning algorithms to identify substance use. Specifically, we adopt several recently proposed techniques to pre-process the text data and propose novel ways of utilizing multi-task learning and multi-modal learning to tackle our problems. The experimental results on the collected data demonstrate the effectiveness of the model, and analyses provides certain insights into our proposed models. We continue to work towards realizing the practical value of our model and to successfully deploy it for substance use prevention in real world settings.

Acknowledgements

Co-author Fang is supported in part by NSF grant IIS-1850477.

[2]https://github.com/AntoinePassemiers/Lexicon-Based-Sentiment-Analysis

Ethical Impact

Our purpose is to examine discrete types of drug use rather than patterns of drug use. It is important to understand what kind of substance homeless youth are using because of the implications and consequences of use. Engagement in some substances such as meth, heroin, and injection drug use are known to have more dire physical and mental health effects than many other commonly used drugs.

Substance abuse is a highly significant public health and social problem in the United States (Tabar et al. 2020; Yadav et al. 2020). While substance abuse is a debilitating problem in its own right, even more importantly, it is a key causative issue for a whole host of other problems faced by homeless youth in their lives, e.g., substance abuse has been shown to increase likelihood of (i) exposure to STIs; (ii) unstable mental health, etc. As a result, it becomes very important to tackle substance use and abuse among homeless youth from a policy planner's and practitioner's perspective. Furthermore, the addictive tendencies associated with substance use mean that it is more cost-effective to prevent substance use before youth get addicted (through proactive interventions) as opposed to treating youth medically after they fall into addiction (through reactive interventions). Social media may offer a powerful opportunity for accessing, educating, and intervening with this typically hard-to-reach group with extremely high rates of drug use. Within this context, we place our work as one of the first attempts at using Facebook data to get weak, low-cost (but hopefully accurate) signals which can be used to identify homeless youth at-risk of substance abuse in the near future.

It is important to consider how the findings of this study can be applied to substance use prevention in real world settings (i.e. non-profit agencies serving homeless youth). One option is to consider engaging Facebook in efforts to use such algorithms to flag users' substance use behaviors. Facebook already uses its own algorithm to detect suicidal ideation. However, such efforts by Facebook have recently become mired in controversy because of concerns with privacy, transparency and ethical issues.

A potential alternative to engaging Facebook would be to create a screening tool that is less likely to violate such privacy and ethical standards and engage agencies that serve this population in utilizing and deploying this tool. For example, most agencies that serve young people who experience homelessness have some kind of an intake process, where youth are screened for various needs and health risks, including substance use. Our algorithmic screening tool can easily be integrated into existing intake processes. These intake processes typically rely on intensive self-reported surveys to screen for substance use. One option is to provide a software tool (or a phone application) that non-profit agencies serving homeless youth can download on their computers/phones which contain (and run) our algorithm. When homeless youth are signing up to receive services at these agencies, they can be asked to volunteer their Facebook conversations along with some simple survey questions to screen them for substance use. To prevent any potential for coercion, youth would be given the option to opt-out of the screening if they had any concerns. However, opting-out would not prevent them from accessing services at that agency. Agencies already have existing protocols that prevent such coercive practices and we can make this opt-out option a part of that protocol. To keep their information safe, their data would be destroyed from the computers once the analysis is run. This would allow agencies to screen young people for substance use without the same privacy and transparency concerns associated with utilization of social media data or the burden associated with intensive surveys. In addition, these agencies may provide online service tools through creating a social media account, and the participants can make some or all of their social media conversations visible to the account. Thus, our algorithm can be run frequently to identify the participants in need of help and support in a timely fashion without asking the participants to fill in lengthy surveys repeatedly.

It should be noted that our proposed system could potentially be misused. For instance, leakage of private Facebook data is a concern, as it means that our system could be used by malicious actors. Also, agencies serving this population might stigmatize youth who are screened as potential drug users and deny them services. Additional efforts are required to prevent the system from being misused.

In addition, it is inevitable that our system could make mistakes and may learn false correlations between Facebook posts and labels. As shown in the analysis section, our model can balance between true and false negative rates by tuning the threshold. Therefore, our model can be adapted to the specific needs of serving agencies. Concretely, if false negative is more costly than false positive, we can set the prediction rate to a high value, and vice versa.

References

Aramaki, E.; Maskawa, S.; and Morita, M. 2011. Twitter catches the flu: detecting influenza epidemics using Twitter. In *EMNLP*, 1568–1576.

Barman-Adhikari, A.; Bowen, E.; Bender, K.; Brown, S.; and Rice, E. 2016. A social capital approach to identifying correlates of perceived social support among homeless youth. In *Child & Youth Care Forum*, 691–708.

Bender, K.; Brown, S. M.; Thompson, S. J.; Ferguson, K. M.; and Langenderfer, L. 2015. Multiple victimizations before and after leaving home associated with PTSD, depression, and substance use disorder among homeless youth. *Child maltreatment* 115–124.

De Lathauwer, L.; De Moor, B.; and Vandewalle, J. 2000. A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications* 1253–1278.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 4171–4186.

Ding, T.; Bickel, W. K.; and Pan, S. 2017. Multi-view unsupervised user feature embedding for social media-based substance use prediction. In *EMNLP*, 2275–2284.

Gerber, M. S. 2014. Predicting crime using Twitter and kernel density estimation. *Decision Support Systems* 115–125.

Guadagno, R. E.; Muscanell, N. L.; and Pollio, D. E. 2013. The homeless use Facebook?! Similarities of social network use between college students and homeless young adults. *Computers in Human Behavior* 86–89.

Jones, S.; and Fox, S. 2009. *Generations online in 2009*. Pew Internet & American Life Project.

Kennedy, D. P.; Wenzel, S. L.; Tucker, J. S.; Green, H. D.; Golinelli, D.; Ryan, G. W.; Beckman, R.; and Zhou, A. 2010. Unprotected sex of homeless women living in Los Angeles County: An investigation of the multiple levels of risk. *AIDS and Behavior* 960–973.

Le, Q.; and Mikolov, T. 2014. Distributed representations of sentences and documents. In *ICML*, 1188–1196.

Marshall, B. D.; Galea, S.; Wood, E.; and Kerr, T. 2011. Injection methamphetamine use is associated with an increased risk of attempted suicide: a prospective cohort study. *Drug and alcohol dependence* 134–137.

Michel, P.; and Neubig, G. 2018. MTNT: A Testbed for Machine Translation of Noisy Text. In *EMNLP*, 543–553.

Nyamathi, A.; Hudson, A.; Greengold, B.; and Leake, B. 2012. Characteristics of homeless youth who use cocaine and methamphetamine. *American journal on addictions* 243–249.

Rice, E.; Barman-Adhikari, A.; Rhoades, H.; Winetrobe, H.; Fulginiti, A.; Astor, R.; Montoya, J.; Plant, A.; and Kordic, T. 2013. Homelessness experiences, sexual orientation, and sexual risk taking among high school students in Los Angeles. *Journal of Adolescent Health* 773–778.

Rice, E.; Milburn, N. G.; and Monro, W. 2011. Social networking technology, social network composition, and reductions in substance use among homeless adolescents. *Prevention Science* 80–88.

Rice, E.; Monro, W.; and Barman-Adhikari. 2010. Internet use, social networking, and HIV/AIDS risk for homeless adolescents. *Journal of Adolescent Health* 610–613.

Sennrich, R.; Haddow, B.; and Birch, A. 2016. Neural machine translation of rare words with subword units. In *EMNLP*, 1715–1725.

Tabar, M.; Park, H.; Winkler, S.; Lee, D.; Barman-Adhikari, A.; and Yadav, A. 2020. Identifying Homeless Youth At-Risk of Substance Use Disorder: Data-Driven Insights for Policymakers. In *KDD*, 3092–3100.

Toro, P. A.; Lesperance, T. M.; and Braciszewski, J. M. 2011. The heterogeneity of homeless youth in America: Examining typologies. *National Alliance to End Homelessness* .

Yadav, A.; Singh, R.; Siapoutis, N.; Barman-Adhikari, A.; and Liang, Y. 2020. Optimal and Non-Discriminative Rehabilitation Program Design for Opioid Addiction Among Homeless Youth. In *IJCAI*, 4389–4395.

Zhou, Y.; Sani, N.; Lee, C.-K.; and Luo, J. 2016. Understanding illicit drug use behaviors by mining social media. *arXiv preprint arXiv:1604.07096* .