A Survey on Illicit Drug Use Patterns by Mining Social Media Data

Vishnu Vikash Medisetty
University of Oklahoma, USA

Le Gruenwald
University of Oklahoma, USA

ABSTRACT

Usage of drugs among people is on the rise and many public health agencies and law enforcement agencies are interested in collecting data on people's drug consumption habits. Illicit drug trading and prescription drug abuse are the major causes behind this. Illicit drug trade via social media sites like Twitter and Instagram has become a severe problem in recent years. Tracking drug dealing and drug abuse on social media sites is of interest to law enforcement agencies and public health agencies. Social media provides a better way to explore the multimedia data, including both images and text in discovering drug use patterns. Using multimodal data on social media enables us in the identification of drug-related posts and analyzing the behavioral patterns of drug-related user accounts. Various methods have been proposed in identifying the illicit drug use patterns by mining social media data. This paper provides a comparative study of techniques used in analyzing the data from these sites and mining them to find common illicit drug patterns and concludes with possible future research directions.

Keywords: Social media, Multimedia, Data Mining, Demographic, Illicit Drug Use, Natural Language Processing, Multimodal Analysis, In-degree, Out-degree, Neural networks, Classifier

INTRODUCTION

In recent years, illicit drug usage and prescription drug abuse has become the major problem for the society and it has been steadily increasing. National Survey on Drug Use and Health (NSUDH) stated that nearly 24.6 million Americans aged 12 or higher were involved in illicit drug usage in a month in the year 2013. It is found that the percentage of total population involved in drug usage is up by 1.1% from 2002 (Xitong & Jiebo, 2017). Also, it is estimated that nearly one third of the people aged 12 or higher involved in drug usage for first time in 2009 have initially started after abusing a nonmedical prescription drug. According to the research reports from National Institute of Drug Abuse, an estimated 48 million Americans (approximately 20% of

the population) aged 12 or higher were involved in using prescription drugs for non-medical reasons at least once in their lifetime. Also, the increase in the number of illicit drug usage is alarming and indicated a growth in illicit drug trade in the US.

With the rapid increase in the internet usage, engaging with social media and access to mobile device among the population there has been an increase in the number of active users. As of 2017, 81 percent of total population (3% percent growth from previous year) in US have a social networking profile there are approximately 2.34 billion social media users worldwide and expected to grow to some 2.95 billion by 2020 (Statista, 2018). National Survey on Drug Use and Health (NSDUH), a national wide survey which aims in collecting data from people on drug usage had 67800 respondents, which is very small compared to number of social media users in US (Yiheng, Numair & Jiebo Luo). Also, these media sites have become effective tools for advertising illegal drugs, with the help of hashtag search a normal person can get access to the information. Due to this few choose to study social media as a data source in understanding drug usage patterns.

The conventional approaches used generally for user data collection are based on recruitment of participants who would provide inputs for a drug-use related study. Some limitations arise for such approaches such as sample size being small, being very costly for involving a large population and mostly the surveys are relied on participant's explicit recall of his/her drug-use behavior. With the vast growth of people using social media and development in the mobile devices there is high possibility of increased drug-use-related data being available online. The user-generated social media collected on a large scale has similar potential of providing insights of drug-usage behaviors, factors and social contexts. (Lee, 2014) has found that substance-use related behaviors have similar patterns when data from social media and traditional survey based approaches are used. (Van Hoof, Bekkers & van Vuuren, 2014) proposed that some Facebook profile elements can be measure of real life behaviors after conducting a study on analyzing Facebook profiles. (Whitehill, Pumper & Moreno, 2015) has conducted study on the relationship between mobile usage of social networking sites and alcohol use in a large street festival. Similarly (Stoddard, S.A., 2012) include the study on influence of young people's social networking behaviors on their alcohol and marijuana usage. From the above work it can be said that social media data can be used in analyzing the drug-usage behaviors.

Most of the work on identifying drug usage and abuse has been done by mining social media data. For example (Yiheng, Numair & Jiebo, 2015) and (Xitong & Jiebo, 2017) are works dedicated in tracking illegal drug dealing and abuse specifically on one social media platform, i.e Instagram. They have used multimodal data and analysis methods such as multi task learning and developing image and text classifiers on posts. Similarly (Cody & Golbeck, 2015), (Carl, Ben, Scott & Christophe, 2013), (Dennis, Melody, Teng-Sheng, 2017),(Ioannis, Azadeh, Matthew, Abeed, Sophia & Gracia, 2016), (Qiongjie, Jashmi, & Baoxin, 2016), (Abeed, Karen, Rachel, & Matthew, 2016) and (Yu & Moh, 2016) are some of the works dedicated in tracking the illicit drug usage patterns using Twitter data as source. Among them (Qiongjie, Jashmi, & Baoxin, 2016) is briefly on using a semi-supervised learning approach to study the behavioral usage of one drug(marijuana). In (Arpita, Anamika, Hamed & Shimei, 2017), they have used Doc2Vec, an unsupervised neural networking-based machine learning algorithm for document embedding as

part of textual data analysis. (Cody & Golbeck, 2015) proposed to analyze the time and location patterns of drug use by mining Twitter data. Similar kind of work can also be found in the (Balasuriya, Wijeratne, Doran & Sheth, 2016) in which Twitter is used to find street gang members and have achieved a good F1 score with classifiers trained on certain features.

Some of the work used Neural Network model for predicting and analyzing the social media data. In (Laura, Diana, Mark, Ethan, Jones & Stephan, 2017), they have compared different machine learning algorithms for identifying the best technique that can predict the success rate of Substance Usage Disorder (SUD) treatment. In (Jaspreet, Mandeep & Gurvinder, 2018), they have used neural network model to analyze the textual data in analyzing the drug addiction causes in Punjab. In (Arpita, Anamika, Hamed & Shimei, 2017), they have used convolutional neural networks in analyzing the image related data and Word2Vec (neural networking approach) in analyzing the textual data. In (Ryan, Deeptanshu & Rahul, 2017), they have used Skip-Gram model (neural networking approach) for computing the semantic relationship between terms and comparing their contexts. In (Priyanka, Sumran and Aleena, 2017), they have used artificial neural networks for prediction of volatile substance abuse. As part of analyzing drug use using social media data one has developed a tool such as PREDOSE (PREscription Drug abuse Online Surveillance and Epidemology) a semantic web platform designed to facilitate the epidemiologic study of prescription drug abuse practices (Delroy, Gary, Raminta, Amit & Drashti, 2013). Some have implemented Map Reduced Model for data mining and storing the bulk amount of data obtained from social media (Shadma, Sonal &Shiv, 2017).

The remainder of this paper is organized as follows: The first section introduces the approaches used in identifying drug-related content on social media. The following section contains a discussion of the main issues that arise while mining drug-related data. Next comes a section that discusses recognition of useful drug related content from data collected from the social media sites and analysis of it. The following section involves literature survey discussing the existing techniques proposed in various papers for analyzing the social media to identify the illicit drug usage patterns. The continuing section involves the summary of data mining techniques applied for analyzing the illicit drug usage to identify the trends. Finally, the last section provides concluding remarks and future research directions.

BACKGROUND

This section discusses the general approach used in identifying drug-related content on social media and presents a discussion about the various techniques used along the process.

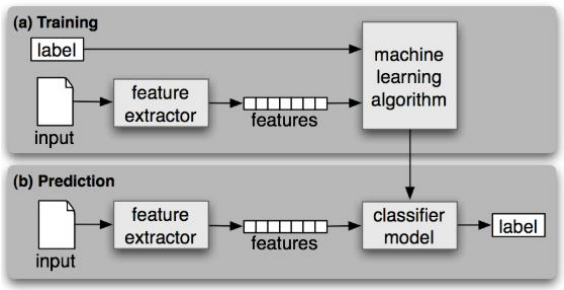


Figure. 1. Supervised Classification

The supervised classification techniques depicted in Figure 1(Steven, Ewan & Edward, 2009) are used in the approach explained below. During the training phase, a feature extractor is useful in generating a feature set from each input value. Basically, the feature sets hold the basic information of each input that can classify it. The labels along with pair of feature sets are fed into the machine learning algorithm to generate a model. During the prediction phase, same feature extractor is used in the conversion of unknown inputs to feature sets. Model is fed with these feature sets, which in turn generates the predicted labels.

The general approach followed by most in the existing papers on identifying drug-related posts by comparing a dynamic hashtag set, and time patterns and mutual interest patterns i.e, drug related posts which are later then analyzed using data-mining algorithms. Multitask Learning is the approach where multiple tasks are learned together thus providing the reader to leverage the information contained in the related tasks which always result to better model for the task. This approach has been used in (Wang, Guo, Lan, Xu & Cheng, 2016) where the authors have characterized the demographic prediction in a retail scenario as a multi-task multi-class problem and then obtained solution by using the structured prediction based on shared representation learning. Similarly, in (Yang & Luo, 2017) the authors use a mask mechanism and a task-relation encoding mechanism into the model. These mechanisms help in training the classifiers for drug related post recognition.

The approach consists of following four steps (Yiang & Luo, 2017). (1) *Potential posts collection*: Collecting a pool of potential drug-related posts using hashtag based-search provided a list of terms related to drug dealing are given to us at the start. (2) *Drug-related post recognition*: Training of Image and Text Classifiers and using them in filtering drug-related posts. Multi task learning is used to

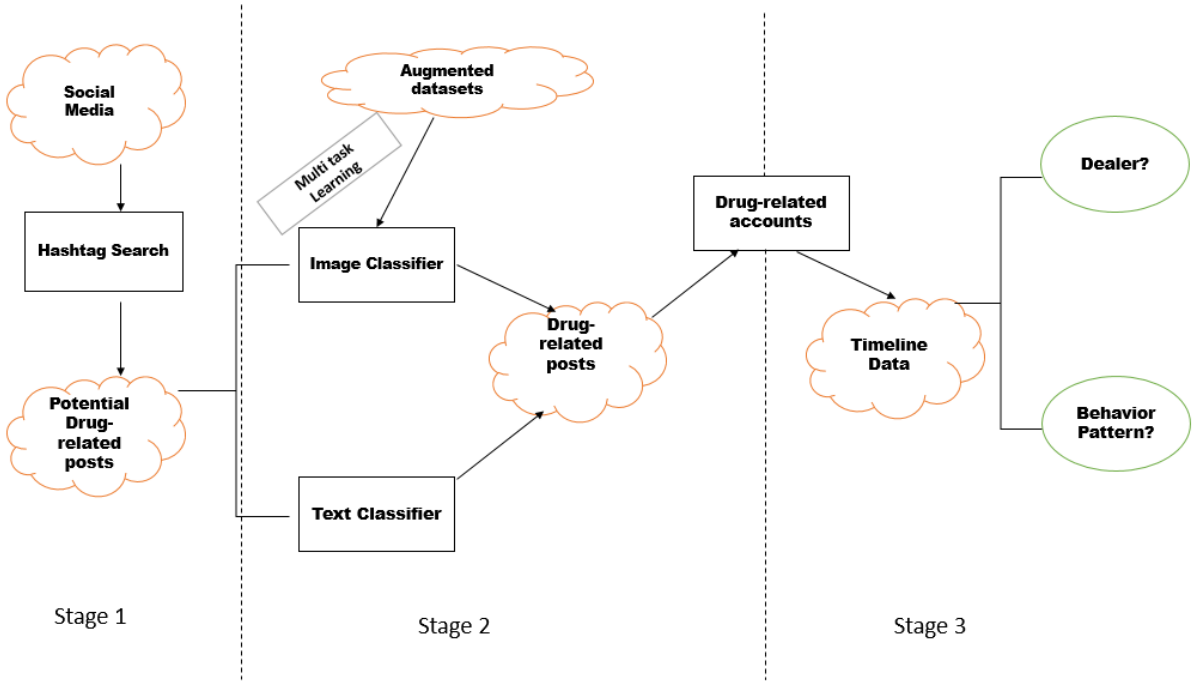


Figure. 2. General framework of the approach used to identify drug related posts on social media sites.

improve the Image Classifier when Images from augmented datasets which are collected from a search engine. Drug related accounts are identified based on the integrated decision made from the Image based and text based classifier. (3) *Account pattern analysis*: Timeline data of drug related accounts are identified and the activity and behavioral related patterns including drug-related patterns, temporal patterns, relational information patterns and so on. (4) *Dealer accounts detection*: Based on set of selected features the drug dealers accounts are determined using a pre-trained classifier.

ISSUES RELATED TO MINING DRUG-RELATED DATA

The following are the issues that should be addressed by the data mining techniques used for analyzing the drug-related data in identifying the usage trends and behavioral patterns:

Slang Terms

Slang terms are major challenging issue for the data mining techniques while analyzing drug-related data. Ambiguity is the major concern as the slang terms have different meaning in different contexts and these terms are mostly used in drug related conversations over the social media sites. Ecstasy slang includes terms like “candy”, “eve”, “molly”, “malcom” and “skittles”. “Skittles” is used as slang term for drug “marijuana”. Similarly, other slang terms generally used for other drugs are “boxes”, “house”, “horse”, “grass”, “glass”, and “pink”. It becomes very difficult to find drug related tweets when such slang terms are used in the posts especially in the case when analysis on the trends and behavioral patterns of the drugs usage is done. There is lot of possibility in obtaining a mixture of relevant and irrelevant content while finding drug-related tweets. In (Cody & Golbeck, 2015), a topic modeling package Mallet (Andrew, 2002) was used to distinguish drug-related topics from others. In (Claudia, Nikki, Thomas, Sean & Adam 2017), they have proposed a vector-space model for predicting relevant drug-related terms like slang terms when the model is provided with a target item (drug-related term). The accuracy of prediction of terms depends on the relevance (>30 %) of term provided as input. However much better topic modelling packages and prediction techniques are needed for tackling this problem.

Evolving drug-related vocabulary

As the illicit drug dealing and spread of drug related information is increasing, the social media sites have started banning the usage of sensitive hashtags or blocking the posts that are associated with such terms. So, the drug-related hashtags have started to evolve to overcome this, everyday new terms (such as “dirtysprite”) are being used in drug-related context in the social media. It becomes difficult to identify drug-related posts with hashtag dataset containing old and no longer used words in drug related conversations. There is a need for constantly updating the hashtag dataset with the current frequently used hashtags. In (Yiheng, Numair & Jiebo, 2016), they have used Apriori algorithm (Agarwal & Srikant, 1994) to generate frequent item sets to update the hashtags dataset.It is required to come up with new approaches in order to tackle this issue more efficiently.

Small Set of Geolocated Data

Even though huge amount of data can be extracted from social media sites using different libraries with the help of the API of medium. But very few posts or twitter feed(tweets) are included with the geolocation data or associated with the global positioning system(GPS) coordinates. So, it becomes difficult in inferring the location attributes of these posts, thus making difficult to analyze the drug usage in different areas. Every tweet or post on Instagram contains metadata field that can optionally include its geographic location and only 2% of the total feed has such information include with them. In (Cody & Golbeck, 2015), they have included this in the future work. It is required to come up with techniques to extract data along with location attributes to have better location based analysis.

Real – Time Analysis

It is generally difficult to acquire recent and timely statistics on drug usage trends and abuse across a wide geographic region (Cody & Golbeck, 2015). Most of the data available for doing analysis to identify the trends comes from surveys released by Substance Abuse and Mental Health Services Administration (SAMHSA) is at least one year old.

Better Classifier

It is difficult to classify drug dealers and drug users from the collection as a drug dealer is also a drug user. In (Yang & Luo, 2017), they have used a predefined blacklist (collection of specific drug-related hashtags) provided by the domain experts in filtering the content by assigning "True" or "False" value to the feature list and using feature list for the classifier in classifying the drug dealer account from the drug user accounts. There is a need for training supervised classifier models on features that better recognize the drug dealers.

Heterogenous Data

It is difficult to classify the heterogenous social media data using a single classifier. Data collected from the social media sites are in the form od text, images and likes (facebook data) and it is difficult to map the features of all types in developing a joint classifier. Decision fusion and multi-task learning are proposed to solve this but the accuracy of the joint models generated was found to be vey less when compared with individual classifiers. In (Xitong & Jiebo, 2017), they tried to implement similar kind of version to some extent and received good results. In (Tim, Warren & Shimei, 2017), they have proposed multi-view user embedding approach to analyze data (likes and posts) collected from Facebook. However, a classifier that can analyze the heterogenous data perfectly is not yet available.

DATA RECOGNITION AND ANALYSIS

In this section we discuss about the recognition of drug-related posts from the data collected over social media. Classifiers must be trained for recognition of drug related posts. First the image – based classifier has to be trained and next text – based classifier has to be trained separately. Then the decisions from both are integrated as weighted average.

Image – Based Classifier

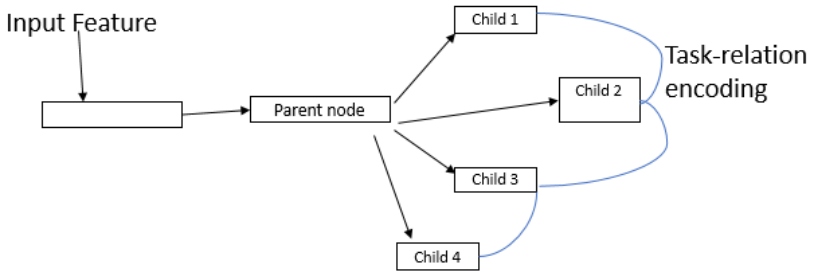


Figure. 3. *Model Architecture for multi-task learning*

Drug-related posts are always in the form of images along with drug-related content in them. It is important to have an efficient image-based classifier to identify drug-related posts. For building a good classifier, large amount of data is required. It is necessary to have high dataset for training the classifier. To have large data for training, data has to be collected from different data domains such as Instagram, Twitter and others from the diverse web data. The datasets from different domains have labels of different meaning. Thus, a proper transfer of information about data and task from the source domain to target domain has to be done. To achieve this multitasking learning method can be used by learning common representations for relevant tasks (Collobert & Weston, 2008). Task- relation encoding is a technique where the task relationship is encoded into structural relationship such that the tasks are represented as a tree structure, parent node being super concept of its children nodes. A parent node is activated i.e labeled as positive when a child node is activated. With the help of this the classifier after seeing a 'marijuana' image will be able to classify it as a drug-related image. Masking mechanism can also be incorporated into the classifier and it allows to ignore the losses caused by other child nodes. In the Figure 3, the four child nodes represent tasks and the parent node is said to be activated if any one of the child nodes is activated, overall it represents the model architecture for multi-task learning. Neural Networks can also be used for analyzing images such as in (Arpita, Anamika, Hamed & Shimei, 2017), they have used convoluted neural networks in the process of class activation mapping for analyzing the important regions in the image.

Text – Based Classifier

Generally, text data includes hashtags and captions is useful in differentiating drug-related and non-drug related posts. Word clouds on the hash tags for both drug-related and non-drug related can be applied to know the most frequent using hashtags in both scenarios. A text based classifier can be trained to recognize drug related posts. Extraction of features for both the hashtags and captions can be done and then features can be scaled by term frequency-inverse document frequency (tf-idf) weighting (Salton & McGill, 1986). Feature dimension can be reduced by retaining the top 1000 features ordered by term frequency across the corpus for tags and captions. A naïve Bayes Classifier (Andrew & Kamal, 1998) can be used for text classification. Neural Networks can also be used in the process of analyzing text such as in (Arpita, Anamika, Hamed & Shimei, 2017), they use Doc2Vec an unsupervised neural network-based machine learning algorithm for analyzing the text.

Fusion of Decision

The decision from both Image and Text Classifier can be integrated to take advantage of the multimodal data on social media. However, it is difficult to have a joint classifier trained on Image and Text data as the correlation between image and data is weaker on the social media. So, training the classifiers separately on their respective data formats and then combining them helps in capturing unique information of different modalities including diverse and subjective behaviors on social media (Xitong & Jiebo, 2017). In (Tao, Warren & Shimei, 2017), they have used mutli-view user embedding (MUE) technique to combine heterogenous data (like and posts).

Behavioral Pattern Analysis

Only small percentage (2%) of total data collected has geographic location included with it in the form of global positioning system (GPS) coordinates. Geolocation information is useful in identifying the places where drug transactions takes place and also helpful in tracking drug dealers. As small percentage of total data collected has such data, it is difficult of perform location-based analysis on the drug-related posts.

The temporal patterns of the drug-related posts can be analyzed with the help of timestamp of each post. Social media posts have timestamp associated with them but some social media sites like Instagram store the GMT time for each post irrespective of the time zone it was generated in. To overcome this, it is feasible to analyze the user accounts associated with these posts and identifying the location of respective user clarifies the time zone of that location. Like this, the time zones can be double checked before analyzing the posts.

The likes and interests of the drug users can be identified by applying the Apriori algorithm to the data to discover the association rules and commonly followed celebrities, common interests and pages. In (Xitong & Jiebo, 2017), they have analyzed the likes, interests and followers of the drug users and found that drug users tend to like certain kind of music and comedians.

LITERATURE SURVEY

This section discusses the various techniques presented in various papers for analyzing the social media to identify the illicit drug usage patterns.

Tracking patterns of activity on Twitter by time and location is growing in fast pace and a wide area of research. By processing the data from social media sites like Twitter, changes in the behavior can be tracked across geographic regions (Conover, 2013). Similarly, (Cody & Golbeck , 2015) have proposed to track changes in the drug usage and popularity of various drugs behavior across geographic regions by processing data from sites like Twitter. This approach is helpful since the National Survey on Drug Use and Health (NSUDH) have grouped many different drug types to large classes as they find difficulty in differentiating them (NSUDH Report, 2014) and analyzing the spread of new drugs. Data is collected from publicly available 1% of the total twitter stream covering from October 30 to November 26, 2014 (81GB compressed). Made use of Apache projects Hadoop, Pig and Spark which leverage the Hadoop file system (HDFS) for distributed processing. Filtering of tweets i.e tweets with no geolocation data is done using twitter library ElephantBird. To overcome the ambiguity of slang terms used in drug context and in normal

context, a topic modelling package Mallet (Andrew, 2002) is used. To determine the popular trends among drugs over time, linear regression was applied to this data and slope of the resulting line is analyzed. Higher slope indicates that the respective drug associated with it has higher popularity

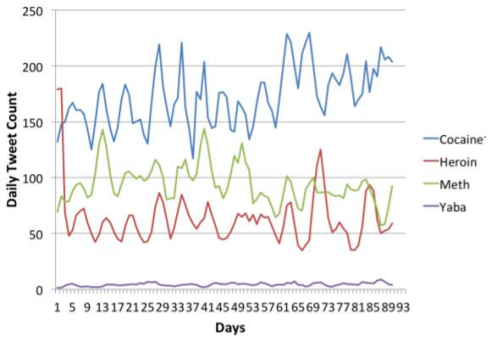


Figure. 4. Daily Tweet Frequencies

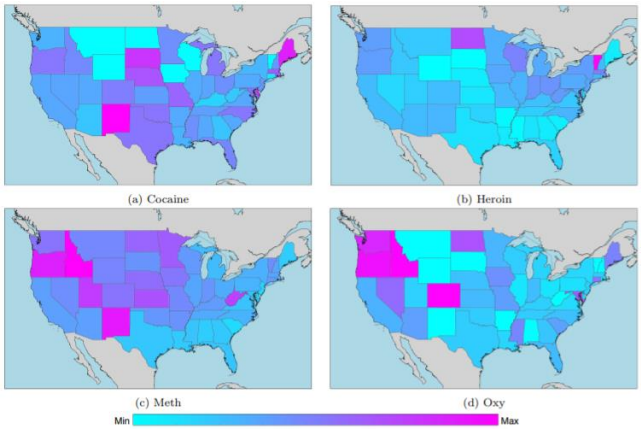


Figure. 5. Geolocated Drug Mentions

In the Figure 4(Cody & Golbeck, 2015) it can be observed that *Cocaine* has higher slope during the days 27-30 and 58 - 63 indicating the increase in the popularity of the drug during those days.

Each tweet is also associated with the geographic location data in the form of global positioning (GPS) coordinates. In general, only two percent of the total tweets contain this geolocated data along with them. In the paper it proposes to identify each state in the US from which each of these tweets were posted. It was achieved by extracting geolocated tweets within United States, dividing them into sets of four classes of drugs and then reverse geo-coding them to identify the US state. Tracking the drug usage like this process would be impractical since states with higher population would produce more tweets thus overshadow the states with lower population that produce higher number of tweets. To overcome this population density affect, normalization techniques are used where each state's tweet count is normalized by the expected number of tweets for that state. Per-state probability distribution Ps(x) = Ns/N, where s { set of US states }, Ns is the number of tweets in state x, and N is the total number of the tweets. The expected number of tweet count in a state is observed to be the product of state's likelihood Ps(x) and total number of tweets. In Figure 5(Cody & Golbeck, 2015), the popularity of drug across geographic locations is illustrated by the per-state distributions with a heat map for each drug in a state.

Some of the issues with the techniques proposed for analyzing temporal and geographic trends for drugs in (Cody & Golbeck, 2015) are disambiguation and small set of available useful data around. With the general usage of slang terms, it becomes difficult to gather all drug -related tweets from the dataset. For example, terms like "weed" are mostly used to describe drug and people use slang terms such as "skittle" in their tweets. Even aggressive filtering will be unable to fully pull drug-related tweets from the data collected over such terms. As only two percent of the total twitter data is included with its geographic location other techniques such as location based analysis approaches can be used.

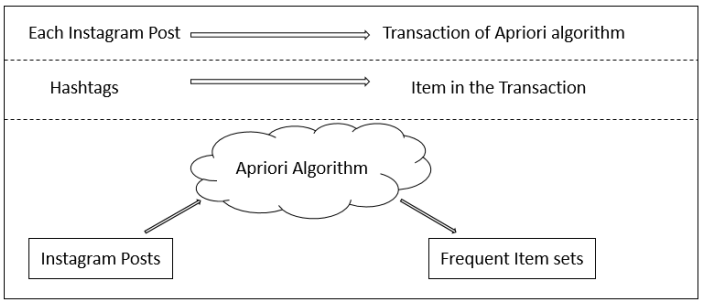


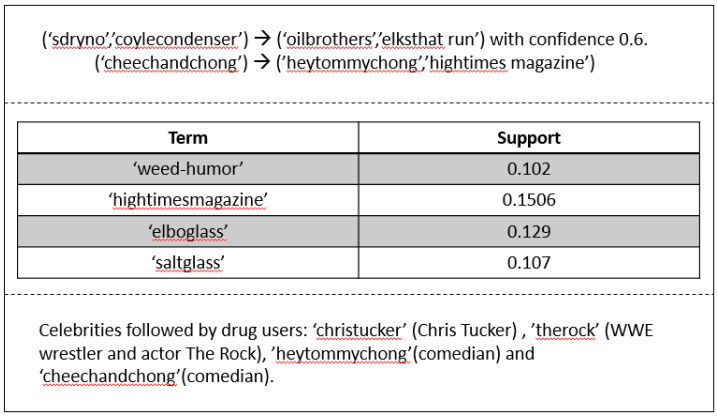
Figure. 6. Generating Frequent Item Sets to keep with up-to-date evolving drug-related hashtags

In (Yiheng, Numair & Jiebo, 2016) fine-grained mining of illicit drug use patters using social multimedia data from Instagram was done. Initially Instagram posts are collected by drug related terms by analyzing the hashtags supplied with each post. They have used PHP's Guzzle library to make request to the Instagram servers to identify and retrieve details of the posts that are drug related. Hashtags attached with the posts are examined to identify the posts contain any sensitive information associated with drug related terms. They have compiled a library of drug related hashtags, manually picked the 100 most frequent drug related hashtags and used tag search API associated with these hashtags. They could get 16 million posts with the help of the API and filtered out the posts that have less than 2 drug-related hashtags. To study the drug consumption patterns, they have searched through all the drug related activity posts to identify the posts that have geolocation tags and applied the Google maps API to identify the locations. They have also used the timestamp associated with the Instagram posts to analyze the frequency of the drug consumption. The timestamps are recorded in GMT time for each post regardless of the time zone it was created in so they have filtered the posts that had matching time zone with the location obtained from the Google Maps API. Drug-related hashtags are evolving as sites like Instagram are banning sensitive hash tags used for drug use. They have used Apriori algorithm (Agarwal & Srikant, 1994) to mine frequently used hash tags to keep up with up-to-date hashtag set. The posts collected with help of initial hash tag set are used in the Apriori algorithm, each post act as transaction of the Apriori and each hashtag as an item in the transaction. Data was stored in CSV files with one transaction each row and hashtags separated by comma. They have applied the Apriori algorithm to those transactions and generated frequent item sets to update the hashtags data set. Figure 6 above depicts the same concept behind generating frequent item sets using Apriori algorithm.These frequent hashtag sets are sorted based on their support, if found support over 20% and if the hashtag is not in the hashtag database then the hashtag is added to the database. As new words are added to the drug vocabulary and these words having different meanings in drug related context it is stated that human supervision is also required to select hashtags before updating the dictionary of frequent drug related terms.

To identify the age and gender information of potential drug users, they have applied the Face API of Project Oxford by Microsoft. They have considered only the posts that are associated with selfie related tags such as 'selfie', 'weedselfie', 'selfpotrait' and 'selfy' in order to calculate the face attributes using Project Oxford. Filtered out posts associated with the above tags but not having image including in it by the help of Detect API of Project Oxford. Increased the accuracy

of age detection by averaging the ages obtained by mining selfies of potential drug users such that each user has 32 selfies in the dataset with a standard deviation of less than 1-year old. With the help of Project Oxford gender detection of the users was also done.

To understand the likes and interests of drug users they tracked the pages a user follows using the Instagram relationship endpoint, recorded them and applied Apriori algorithm to the data for discovering commonly followed pages and association rules. They stated that pages with high supports and association rules that had a probability of 1 were found. Some of the association rules from (Yiheng, Numair & Jiebo, 2016) are: (‘sdryno’,’coylecondenser’) → (‘oilbrothers’,’elksthat run’) with confidence 0.6 and (‘cheechandchong’) → (’heytommychong’,’hightimes magazine’) where ‘elksthatrun’ is a music band indicating that drug users generally like certain kind of music and ‘cheechandchong’ is a comedian, people who like him also like ‘hightimesmagazine’ indicating drug users following some particular comedians. It is also stated that drug users follow glassmakers such as ‘elboglass’ and ‘slatglass’ with support 0.129 and 0.107 values. It is stated


(‘sdryno’,’coylecondenser’) → (‘oilbrothers’,’elksthat run’) with confidence 0.6.
(‘cheechandchong’) → (’heytommychong’,’hightimes magazine’)

Term	Support
‘weed-humor’	0.102
‘hightimesmagazine’	0.1506
‘elboglass’	0.129
‘saltglass’	0.107

Celebrities followed by drug users: ‘christucker’ (Chris Tucker) , ’therock’ (WWE wrestler and actor The Rock), ‘heytommychong’(comedian) and ‘cheechandchong’(comedian).

Figure. 7. Likes and Interests of Drug Users

that other interests like celebrities, people, music, magazines, music and all things that drug users like and follow can be identified.

The social networks of drug users and non-drug users were also compared to identify the differences. Non-drug related users are collected by using a simple classifier on the posts of the users obtained by filtering by using non-drug related hashtag. The classifier works in such a way that it searches over the all the hashtags associated in the users posts and the user is eliminated from the non-drug user’s data if drug-related hash-tag was found in his posts. They created two graph objects out of the data one for drug-related and non-drug related accounts. Computed various statistical measures on the two graph objects and it was stated that examining the vertices with maximum in-degree in each of the networks, a clear difference in the preferences of two groups were noted such as non-drug users frequently liked pages about fashion and makeup and drug users like pages that are about drugs. It is also stated that examining the 3 node cycles of the graph and identifying the vertex involved in the most node cycles in both the graphs, found that in drug-related graph it was a drug dealer was most common involved vertex and in the non-drug related it was a popular user on Instagram. Drug user network had ten times more cycles than non-drug user networks (Yiheng, Numair & Jiebo, 2016) indicating that drug users form more connected

and centralized community than non-drug related users. Also proposed that drug dealers have relatively low in-degree and out-degree compared with drug users and non-drug users but drug dealers have higher ratio of in/out- degree than regular drug users indicating that dealers are more important vertices that connect drug user's community in network than regular drug users.

In (Xitong, Jiebo, 2017), they have proposed a framework for identifying drug-related posts, analyzing behavioral patterns and detecting drug dealer account information to some extent on Instagram. Multimodal data and analysis methods such as employing multi task learning approach for Image based classifier and text based classifier for identifying drug related images posts and hashtags. Also used the task relation-encoding approach and masking mechanism for creating a better Image classifier. For text-based classifier approaches like feature extraction, scaling by term frequency-inverse document frequency reducing feature dimension by retaining top features and using a naïve Bayes classifier at the last. The fusion of results from both text based classifier and image based classifier was done as training joint model classifier is challenging and ineffective and training two separate classifiers can help in better capturing of unique information of different modalities. Also proposed to detect drug dealer account information with the help of evidence of transactions approach, a technique where the data extracted on bio description and comments of the drug-related posts is filtered by a predefined blacklist provided by the domain experts. True value is assigned if one of the blacklist terms occur. They have also trained a linear logistic regression classifier with L1 regularization on hold-out data and feature dimension values with zero or very small coefficient values were removed.

There has also been work done on developing semi-supervised model for studying the user behaviors of illicit drug using social media data. One such is (Qiongjie, Jashmi & Baoxin, 2016), they proposed a semi-supervised approach to study the illicit marijuana usage using noisy, unstructured and large-scale Twitter data. The approach followed was to initially extract a set of basic features from each tweet and then utilizing a small labeled training set, a good feature mapping is learnt that takes into consideration both the basic features and their interactions based on weakly-hierarchical lasso. Finally, the learned feature mapping model was used to process the large-scale Twitter data. The features extracted in initial stages are of two types: Content based features involve length of tweet, favorite count & retweeted count, number of hash tags and TF-IDF on Unigram, and user-based features involve number of followers and followings and number of tweets.

In (Jaspreet, Mandeep & Gurvinder, 2018), they have evaluated the causes of drug addiction in Punjab and its effect on life using machine learning techniques. They have proposed a model which involves neural network based classification method implemented in R language for classification of text reported for drug abuse available on social media sites and web sources. Data preprocessing and labelling of training data set was done manually by experts. The proposed neural network model was trained using this data and a word cloud was generated. Word cloud was compared with collection of positive and negative word across the globe and more negative words were matched with the global collection than negative words indicating that drugs has serious effect on the life of people. Accuracy of the model was predicted by testing on the test data was found to be highest (84.35%) compared to other works (70%) done on the same topic.

In (Arpita, Anamika, Hamed & Shimei, 2017), they have proposed an automated detection of substance-use related social media posts based on image and text analysis approaches. Data was collected from Instagram site using a bootstrapping process for social media post retrieval and hashtag-based search. They have used a set of hash tags which includes official and street name of drugs found on National Institute on Drug Abuse (NIDA) website. They have used convolutional neural network technique for learning image based features and neural network classifier that uses Doc2Vec for learning text based features. Doc2Vec being a neural network machine learning algorithm that learns fixed-length feature representations from variable length texts(sentences), they have found it very useful in analyzing text based features. They have combined features learned from both techniques and tried to find the best model that can automatically predict illicit drug-related posts. The best model they achieved had 90% accuracy and 75% F1-measure, whose performance was significantly better than models that used only one type of features.

In (Carlo & Rada, 2017), they have proposed a computational analysis of drug addict experiences in order to understand the characteristics of drug users. The data related to drug experiences was collected from www.erowid.org website that included 4 types of drugs categorized according to their main effects. They have used a Multinomial Naïve Bayes classifier with information gain feature weighting this approach for classifying the documents. They have analyzed the distribution of psycholinguistic word classes to quantify the similarity of distributions and emotional experiences of psycholinguistic processes across the four drug types. By calculating the Pearson correlation between the dominance scores of linguistic classes they have found drug types that are similar to other types and also emotional experiences associated with each drug type.

In (Ravneet, 2016), the author proposed to predict the behavior of drug addiction with the help of Fuzzy Expert System. Implemented an artificial intelligence approach that involves gathering of membership functions and rules to analyze the features of the data. Fuzzy verdict mechanism is implemented to analyze the factors selected from the data collected and prediction is done based on the fuzzy expert system. They have been able to predict the behavior of drug addiction with an accuracy of 70%.

In (Tim, Jannani, Takeo & Gert, 2017), they have proposed to detect illegal online sale of prescription opioid. The process involved identifying and characterizing the illicit online sale of controlled substances via twitter. They have used Amazon Web Services EC2 t2.micro virtual instances that are preconfigured with RStudio to collect large volumes of tweets filtered by the opioid keyword set from twitter. For isolating the word groupings associated with tweets that mentioned marketing and sale of prescription opioid drugs they have used Biterm Topic Model (BTM), an unsupervised machine learning approach was used. Initially preconfigured themes and patterns detected by BTM were then coded resulted in achieving high intercoder reliabilities for BTM word grouping inclusion criteria. Content analysis was performed to identify type of web site, legal status and domain registration information. They have found that number of tweets related to illegal promotion and sale were less when compared to the tweets on behavioral aspects of drug usage.

In (Priyanka, Sumran and Aleena, 2017), they have proposed a machine learning approach to predict volatile substance abuse for dug risk analysis. They have used two artificial neural networks in the prediction of volatile substance abuse. ANN-D was used to predict whether a person is volatile substance abuse user or not. If found to be volatile substance abuse user then ANN-C was used to predict the time (day, week, month, year, decade or before a decade) of use. The input features used are age, gender, country, ethnicity, education, neuroticism, openness to experience, extraversion, agreeableness, conscientiousness, impulsiveness, sensation seeking etc. The accuracy of ANN-D and ANN-C was found to 81% and 71.9% respectively, the results are obtained after performing 10-fold cross validation.

In (Ryan, Deeptanshu & Rahul, 2017), they have proposed a model to identify individuals amenable to drug recovery interventions through computational analysis of addiction content in social media. Data was collected from the recreational drug use and addiction recovery forums on reddit website. For generating dynamic lexicon, a Density Based Lexicon Expansion with Seeding (DBLES) was used, which involves the application of Skip-Gram model (a single layer neural network approach) to compute the semantic relationships between terms. For identifying the drug-related subreddit, a biclustering (Kluger, 2003) approach and its results were used. For predicting the addiction propensity, k-NN classifier with k=11 is used as it provided best results with an F1 score of 0.848. They found helpful in understanding the factors that drive and influence addiction with these approaches.

In (Nhathai, Soon, Manasi & James, 2017), they have demonstrated the possibility of utilizing social media, for automatic monitoring of illegal drug and prescription medication abuse. Data associated with the set of well-known illegal and prescription drugs according to National Center on Addiction & Substance Abuse, was collected from twitter. Slang terms were also included in the above set to get better results. They have manually classified the tweets into drug abuse tweets and non-abuse tweets. Tableau was used for data analysis and visualization, it also provided them insights on correlation among user behaviors and drugs. Based on these insights they have used String2WordVector algorithm which converts string attributes of tweets into set of attributes representing word occurrences based on TF/IDF. These word occurrences were used in Random Forest, Decision Tree(J48), SVM and Naïve Bayes machine learning models. All these models are evaluated by applying them to 300 manually labelled tweets with 10-fold cross-validation (training-set =67%). J48 model achieved high precision of 74.8% among them.

In (Tao, Warren & Shimei, 2017), they have showed that effective social media based substance use detection systems can be built using machine learning and text mining techniques which can be used to identify people who are at the risk of SUD (Subtance Use Disorder). Data was collected from Facebook application and they obtained a very huge unsupervised like dataset. Unsupervised feature learning techniques were used to learn a dense vector representation of a user's Facebook posts and likes to take advantage of large amount of unsupervised data. Single-view Post Embedding (SPE), Single-view Like Embedding(SLE) and Multi-view user Embedding (MUE) were proposed for Facebook's posts, likes and combination of likes and posts data (heterogenous). The Feature learning methods for SPE are Singular Value Decomposition, Latent

Dirichlet Allocation, Document Embedding with Distributed Memory (D-DM) and Document Embedding with Distributed Bag of words (D-DBOW). Among them User Document Embedding with Distributed Bag of words (D-DBOW) was best for SUD prediction. The Feature learning methods for SLE are Singular Value Decomposition, Latent Dirichlet Allocation, Document Embedding with Distributed Memory (D-DM), Document Embedding with Distributed Bag of words (D-DBOW) and Auto Encoder(AE). Among them Document Embedding with Distributed Bag of words (D-DBOW) was best for SUD prediction. Features learning methods for MUE are: Canonical Correlation Analysis (CCA) and Deep Canonical Correlation Analysis (DCCA). Among them wGCCA_balanced and wGCCA_imbalanced were best for (Alcohol & Drug) and Tobacco SUD prediction.

In (Ding, Roy, Chen, Zhu & Pan, 2016), they have applied topic modelling to disambiguate hashtags and track the changes of hashtags using semantic word embedding. Latent Dirichlet Allocation, a topic modelling technique used to discover main topics in a text collection and K-means clustering to group posts with similar topic together. With T (parameter of LDA) and K(K-means) set to 30 and 5 respectively, 78.1% accuracy was achieved in identifying illicit drug-related posts on social media. For the detection of new illicit drug related hashtags, a neural word embedding technique with help of skip-gram model (a neural networking approach) to predict words in the target context, was used to get best results.

In (Abeed, Karen, Rachel, & Matthew, 2016), they have designed an automatic supervised classification technique to distinguish posts containing signals of medication abuse from those that do not and assessed the utility of twitter in investigating patterns of abuse over time. Data related to three abuse prone medications (Adderall, OxyContin & Seroquel) based on their abuse potential was collected from twitter. They have considered six feature sets such as word-n-grams, abuse-indicating terms, drug-slang lexicon, synonym expansion, word clusters and classification. For classification four classifiers are considered Naïve Bayes, Support Vector Machines (SVMs), Maximum Entropy(ME), and a decision tree-based classifier(J48) were considered. Stacking based classifier is proposed where the predictions from all classifiers are combined and another classifier is trained to make final decision based on individual predictions. It is found that Stacking Classifier developed achieved highest F1 scores. It is found that when 10000+ annotated training instances were used, F1 score of more than 0.55 can be achieved. They have showed the temporal patterns of abuse indicating tweets for each of the drug. They have also conducted single-feature and leave-out feature experiments to study the impact of each feature in the classification. They showed that social networks like Twitter contains valuable information in terms of medical abuse.

In (Andrei, Y. & Sergey, M., 2014), they have proposed a solution for mining and analysis of data from the social media. Data mining from the social media was done by using crawler based on MapReduce model for distributed computations and implemented using Hadoop framework. They have use weighted key phrase technique, where weights of key phrase were made larger than the sum of keywords weight indicating key phrase as stronger signal of text relevance to drug theme. Documents with largest weights were assumed to be more relevant to a drug theme and if users include any of the documents, they were assumed to be interested in drug theme. Using this approach, they have identified the interest levels of drug users in social themes and compared them

with non-drug users. They have generated a model with subgroups such as age and sex along with factors such as unemployment level, gini index, ratio of morality and fertility rates, ratio of divorces and ages and life satisfaction. The evolution of each group was analyzed as Markov Chain problem and Pierson correlation was done between the groups and factors that influence on level of drug usage. The data analysis process was formalized in application that was run on distributed computing – based cloud platform (CLAVIRE – Cloud Applications VIRtual Environment) environment.

In (Delroy, Gary, Raminta, Amit & Drashti, 2013), they have discussed about semantic web platform called PREDOSE (PREscription Drug abuse Online Surveillance Epidemiology) they designed to facilitate the epidemiologic study of prescription drug abuse practices using social media. They have collected data from three online forums, modelled in a manually created Drug Abuse Ontology (DAO) and used a combination of lexical and semantic techniques to facilitate the extraction of semantic information (entities and relationships) from the downloaded data. They have used probabilistic optimization algorithm to extract the final sentiment expressions which help to capture of fine-grained semantic information that facilitate search, trend and overall content analysis. They evaluated the extraction techniques that indicated 85% precision and 72% recall in entity identification, 36% precision in relationship identification and 33% precision in triple extraction.

In (Laura, Diana, Mark, Ethan, Jones & Stephan, 2017), they have used machine learning framework to predict the success rate of substance use disorder treatment. They have also used new machine learning method i.e. Super Learning (SL), an ensemble machine learning method which is a generalization of stacking algorithm (Wolpert, 1992) that takes weighted average of all the algorithms considered for prediction and produces a single prediction function (PF). They have used 28 predictors obtained from the Treatment Episode Data Set - Discharges(TEDS-D) as a predictor set and associated along with binary outcome treatment success (Yes/No) to generate random variable (O) upon which true probability distribution P_0 was calculated. They estimated the probability of succeeding in a treatment(Q_0), $Q_0 = P_0(Y=Yes|W)$ and tried to find the machine learning algorithm that finds the best estimate of Q_0, which was achieved by maximizing AUC. To avoid overfitting, they have used 2-fold cross-validation (CV) and model was validated on randomly selected data. The machine learning algorithms such as Logistic regression, Penalized regression, Random Forests, Artificial Neural Networks and Super Learning were compared for the set of predictors and analyzed the AUC values. Their results showed that all the algorithms obtained AUC value between 0.793 and 0.820 indicating that these algorithms have higher probability of randomly choosing successful patient over an unsuccessful patient. SL has the highest AUC value followed by Random Forests among the others.

In (Claudia, Nikki, Thomas, Sean & Adam 2017), they have predicted drug-terms relevant to target terms on the data collected over social media. They have used vector-space model to encode the semantic relationships from spatial relationships and was trained on the data collected from the Twitter site. They have associated cosine similarity to the models for better results in detecting drug-abuse related terms relevant to the target drug-terms provided to the model. They

showed that model provided better results in predicting drug-terms that are more than 30% relevant to the target drug-term provided as input over low relevant drug terms.

In (Shadma, Sonal &Shiv, 2017), they have proposed a tool for managing bulk amount of data with mining in social media on composite applications for performing complex analysis using cloud platform. They have indicated the usage of Big Data handling techniques such as Map Reduce for mining and analyzing social media data. Data from the social media sites was mined using crawler (Glotzer, 2009) and saves the data into Hadoop cluster. Data collected was filtered and aggregated in order to get task related data. They have proposed to use classification technique to predict the drug usage behavior. The collected data was provided as input for complex applications that are managed by model implemented in distributed computing-based platform environment, that perform data analysis.

SUMMARY OF TECHNIQUES

The summary of algorithmic techniques employed in the analysis of illicit drug use:
In the table below, A denotes Algorithm used in the analysis (S - Supervised, SS - Semi-Supervised, US - UnSupervised), I – Dataset source, II - Image data analysis, III - Text data analysis, IV - Multimodal analysis, V - Drug popularity analysis, VI - Location based patterns, VII - Temporal patterns, VIII - Social network analysis and IX - Behavioral traits analysis.

Sr. No	Paper	Technique	A	I	II	III	IV	V	VI	VII	VIII	IX
1.	(Tim, Janani, Takeo & Gert, 2017)	Biterm Topic Model along with Content Analysis	US	Twitter	×	✓	×	×	✓	×	×	×
2.	(Qiongjie ,Jashmi & Baoxin, 2016)	Compared linear Support Vector Machines (SVM), Linear Classifier (LC))	SS	Twitter	×	✓	×	×	×	×	✓	✓
3.	(Arpita, Anamika, Hamed & Shimei, 2017)	Convolutional Neural Network Model for Image based features and Doc2Vec algorithm for text based features	SS	Instagram	✓	✓	✓	×	×	×	✓	×

4.	(Xitong & Jiebo, 2017)	Neural Network model for Image analysis and Naïve Bayes Classifier for text analysis	S	Instagram	✓	✓	✓	×	×	✓	✓	✓
5.	(Jaspreet, Mandeep & Gurvinder, 2018)	Neural Network model on the textual data	S	National Household of survey of Drug abuse(NHSDA)	×	✓	×	×	×	×	×	✓
6.	(Priyanka, Sumran & Aleena, 2017)	Artificial Neural Network (ANN) used for prediction of volatile substance abuse. ANN-D to predict drug user or not and ANN-C to predict the time of last drug use	S	UCI Machine Learning Repository	×	✓	×	✓	×	✓	✓	✓
7.	(Ryan, Deeptanshu & Rahul, 2017)	Skip-Gram, a neural network model for computing semantic relationships and k-NN classifier to predict addiction propensity	S	Reddit	×	✓	×	✓	×	×	✓	✓
8.	(Andrei & Sergey, 2014)	Weighted key phrase technique used to analyze the text and Model associated with Pierson correlation and Markov chain techniques to analyze the behavioral aspects of addicts.	S	Social Media Sites	×	✓	×	✓	✓	×	✓	✓

9.	Shadma, Sonal & Shiv, 2017)	Map Reduce Model used to mine and process social media data. Classification technique proposed for prediction.	S	Social Media Sites	×	✓	×	×	×	×	×	×
10.	(Nhathai, Manasi, Soon & James 2017)	Artificial Neural Network (ANN) used for prediction of volatile substance abuse. ANN-D to predict drug user or not and ANN-C to predict the time of last drug use	S	Twitter	×	✓	×	✓	×	×	✓	✓
11.	(Abeed, Karen, Rachel, & Matthew, 2016.)	Naïve Bayes, Weighted Support Vector Machine (SVM), Maximum Entropy, Decision-Tree(J48) to assess the performance of automatic detection	S	Twitter	×	✓	×	✓	×	✓	×	×
12.	Laura, Diana, Mark, Ethan, Jones & (Tim, Janani, Takeo & Gert, 2017)	Compared Logistic regression, Penalized regression, Random Forests, Artificial Neural Networks, Super Learning (ensemble machine learning method)	S	Treatment Episode Data Set-Discharges(TEDS-D), 2006-2011 data	×	✓	×	✓	×	×	×	×

		for classifying text.										
13.	(Claudia, Nikki, Thomas, Sean & Adam 2017)	Vector Space Models were used to encode semantic relationships	S	Twitter	×	✓	×	✓	×	×	×	×
14.	(Qiongjie, Jashmi & Baoxin, 2016)	Compared linear Support Vector Machines (SVM), Linear Classifier (LC) with proposed semi-supervised learning approach	S	Twitter	×	✓	×	✓	×	×	×	×
15.	(Delroy, Gary, Raminta, Amith & Drashti, 2013)	Combination of lexical & semantic based techniques to extract entities and relationships. Probabilistic optimization algorithm to extract sentiment expressions.	S	Social media sites	×	✓	×	✓	×	✓	×	✓
16.	(Ravneet Kaur, 2017)	Fuzzy Expert System associated with membership functions and rules are used for analyzing textual data	S	Survey based dataset	×	✓	×	×	×	×	×	✓
17.	(Ding, Roy, Chen, Zhu & Pan, 2016)	LDA and K-means clustering for identifying drug-related data & Skip-Gram model (neural network approach) for	S	Instagram	×	✓	×	✓	×	×	×	×

		training word-embeddings used in prediction of new hashtags.										
18.	(Yiheng, Numair & Jiebo, 2016)	Apriori algorithm for frequent item-sets generation to Update hashtags dataset	S	Instagram	✓	✓	✓	✓	×	✓	✓	✓
19.	(Cody & Jeniffer, 2015)	Linear Regression and analyzed slope and squared Pearson Coefficients (R_2)	S	Twitter	×	✓	×	✓	✓	✓	×	×
20.	(Pollard & Homan, 2016)	SVM with linear kernel using F-1 score	S	Twitter	×	✓	×	✓	×	×	×	×
21.	(Carlo & Rada, 2017)	Multinomial Naïve Bayes Classifier using the Information Gain for classifying the documents	S	www.erowid.com (web source)	✓	✓	✓	✓	×	×	×	✓

CONCLUSION AND FUTURE RESEARCH DIRECTIONS

This work identified the data mining approaches and research scope issues concerning the analysis of illicit drug use trends. (Cody & Jennifer, 2015) have difficulty in identifying drug - related tweets from the data collected due to the presence of slang terms. (Yiheng, Numair & Jiebo, 2016) proposed approach is so far, the best approach among the available techniques in analyzing

the drug-related trends. It was the first to analyze the social network of the drug-related users and also the first to analyze the behavioral traits of drug-users such as their interests and pages that they follow. It also suggests a way to identify the drug dealers from the drug users. The paper has to come up with new approach in terms of location based analysis as very few data have included with geolocation information among the large amount of data collected. Also (Xitong & Jiebo, 2017) and (Arpita, Anamika, Hamed & Shimei, 2017) proposed approaches to classify both image based and text based data easily. (Xitong & Jiebo, 2017) proposes a new way to detect drug dealer account information but it fails to solve the problem behind location based analysis and behavior trait analysis. (Qiongjie, Jashmi & Baoxin, 2016) is the first semi supervised approach used to analyze the illicit drug usage. The approach should be improved such that using this approach other feature analysis such as location based analysis, behavioral traits analysis and network analysis of drug-related terms can be done easily. (Pollard & Homan, 2016) is not a fully developed approach, the classifier must be trained on large data to get better results and with further tuning of the classifier location based, behavioral traits and network analysis can be performed. In (Shadma, Sonal &Shiv, 2017), they have used map reduce model for mining, storing and processing of the bulk data collected from social media sites, however they have not discussed the classification approaches for predicting the data. It is found that along with classifiers people have used neural network techniques in analyzing the image and textual data. Analyzing the heterogenous data with a joint classifier for both types of data is still a concern. Better topic modelling packages like Mallet must be developed as they are helpful in differentiating drug-related terms from non-drug related terms when slang terms are present in the posts. As Illicit drug usage is on rise and has become major problem for the society. Law enforcement agencies and health agencies are interested in tracking the illicit drug use and abuse. By developing more robust techniques in analyzing the data and providing more support to research in tracking the illicit drug usage and understanding the trends, the drug abuse and overdose can be minimized.

REFERNCES

Abeed, S., Karen, O., Rachel, G. & Matthew, S., Smith, K., Dan, M. & Graciela, G. (2016). Social Media Mining for Toxicovigilance: Automatic Monitoring of Prescription Medication Abuse from Twitter. *Drug Safety, vol. 39, no.3*

Agrawal, R. & Srikant, R. (1994). Fast algorithms for mining association rules in large databases. *Proceedings of the 20th International Conference on Very Large Data Bases, pages 487-499.*

Agarwal, A., Xie, B., Vovsha, I., Rambow, O. & Passonneau, R. (2011). Sentiment analysis of Twitter data. *In Proceedings of the Workshop on Languages in Social Media (LSM '11).*

Andrei, Y. & Sergey, M. (2014). Social Networks mining for analysis and modelling drugs usage. *14th International Conference on Computational Science.*

Andrew, K.M. (2002). Mallet: A machine learning for language toolkit. *http://mallet.cs.umass.edu/*

Andrew, M. & Kamal, N. (1998). A comparison of event models for naïve bays text classification. In *Proceedings of AAAI-98 Workshop on Learning for Text Categorization, Vol.752.Citeseer, 41-48.*

Arpita, R., Anamika, P., Hamed, P. & Shimei, P. (2017). Automated Detection of Substance Use-Related Social Media Posts Based on Image and Text Analytics. *International Conference on Tools with Artificial Intelligence.*

Balasuriya, L., Wijeratne, S., Doran, D. & Seth., A. (2016). Finding Street gang members on twitter. *ACM International Conference on Advances in Social Networks Analysis and Mining*.

Carl, L.H., Ben, C., Scott, B. & Christophe, G.C. (2013). An Exploration of Social Circles and Prescription Drug Abuse Through Twitter. *Journal of Medical Internet Research.*

Carlo, S. & Rada, M. (2017). A Computational Analysis of the Language of Drug Addiction. *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL.*

Claudia, B., Nikki, A., Thomas, C., Sean, S. & Adam, L. (2017). Drug term discovery through social media using natural language processing. *National Drug Early Warning System, University of Maryland.*

Cody, B. & Jennifer, G. (2015). This is your twitter on drugs: Any questions? *International Conference on World Wide Web Companion, pages 777-782.*

Collobert, R. & Weston, J. (2008). A unified architecture for natural language processing: Deep neural network analysis of Instagram user timelines. *In proceedings of the 25th International Conference on Machine Learning, ACM, 160-167.*

Conover, D.M., Davis C., Ferrara, E., McKelvey K., Menczer F. & Flammini A. (2013). The geospatial characteristics of a social movement communication network. *PloS one, 8(3):e55957.*

Dennis, H., Melody, M. & Teng-Sheng, M. (2017). Mining Frequency of Drug Side Effects over a Large Twitter Dataset Using Apache Spark. *ACM International Conference on Advances in Social Networks Analysis and Mining*.

Delroy, C., Gary, A.S., Raminta, D., Amit, P.S., Drashti, D. (2013). PREDOSE: A semantic web platform for drug abuse epidemiology using social media. *Journal of Biomedical Informatics.*

Ding, T., Roy, A., Chen, Z., Zhu, Q. & Pan, S. (2016). Analyzing and Retrieving Illicit Drug-Related Posts from Social Media. *IEEE Conference on Bioinformatics and Biomedicine.*

Glotzer, S.C. (2009). WTEC Panel Report on international assessment of research and development in simulation based engineering and science. *World Technology Evaluation Centre.*

Ioannis, K., Azadeh, N., Matthew, S., Abeed, S., Sophia, A. & Graciela, H.G. (2016). Analysis of the effect of sentimental analysis on extracting adverse drug reactions from tweets and forum posts. *Journal of Biomedical Informatics.*

Jaspreet, S., Mandeep, K. & Gurvinder, S. (2018). Evaluation of Drug Addiction Causes in Punjab using Machine Learning. *International Journal of Advanced Research in Computer Science*

Kluger, Y. (2003). Spectral biclustering of microarray data: coclustering genes and conditions. *Genome research 13.4, 703-716.*

Laura, A., Diana, K., Mark, V.D.L., Ethan, S., DeShauna, J. & Stephen A. (2017). Use of a machine learning framework to predict substance use disorder treatment success. *PLOS ONE*

Lee, C. (2014). Recruitment through social networking sites: Are substance use patterns comparable to traditional recruitment methods? *Journal of Medical Internet Research.*

Liang, W., Teng-Sheng, M. & Natalia, K. (2015). Twitter opinion mining for adverse drug reactions. *Proceedings of the 2015 IEEE International Conference on Big Data (Big Data).*

Nhathai, P., Manasi, B., Soon, A.C. & James, G. (2017). Enabling Real-Time Drug Abuse Detection in Tweets. *IEEE 33rd International Conference on Data Engineering.*

Peng, Y., Moh. M., & Moh. T. (2016). Efficient Adverse Drug Event Extraction Using Twitter Sentiment Analysis. *Proceedings of the 8th IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM).*

Pengfei, W., Jiafeng, G., Yanyan, L., Jun, X. & Xueqi, C. (2016). Your cart tells you: Inferring demographic attributes from purchase data. *In proceedings of the Ninth ACM International Conference on Web Search and Data Mining.*

Pollard, D. & Homan, C. (2016). Detecting Illicit Drug Usage on Twitter. *semanticscholar.org, 06f9*

Priyanka, N., Sumran, K. & Aleena, S. (2017). A Machine Learning Approach to Predict Volatile Substance Abuse for Drug Risk Analysis. *3rd International Conference on Research in Computational Intelligence and Communication Networks(ICRCICN).*

Qiongjie, T., Jashmi, L. & Baoxin, L. (2016). Finding Needles of Interested Tweets in the Haystack of Twitter Network. *ACM International Conference on Advances in Social Network Analysis and Mining.*

Ravneet, K. (2016). Prediction of Drug Addiction in Punjab with Fuzzy Expert System. *International Journal of Engineering Science and Computing.*

Ryan, E., Deeptanshu, J. & Rahul, S. (2017). Identifying Individuals Amenable to Drug Recovery Interventions through Computational Analysis of Addiction Content in Social Media. *IEEE International Conference on Bioinformatics and Biomedicine(BIBM).*

Salton, G. & McGill, J.M. (1986). Introduction to Modern Information Retrieval. *Mc-Graw-Hill, Inc. New York, NY.*

Shadma, Q., Sonal, R. & Shiv, K. (2017). Mining Social Media Data for Understanding Drugs Usage. *International Research Journal of Engineering and Technology(IRJET).*

Steven, B.,Ewan, K. & Edward, L. (2009) Learning to Classify Text. *Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit, ed. 1, ch. 6.*

Statista (2018). Social media- Statistics & Facts.

https://www.statista.com/topics/1164/social-networks/

Stoddard, S.A., Bauermeister, D., Messer, G., Johns, M. & Zimmerman, M.A. (2012). Permissive norms and young adults alcohol and marijuana use: The role of online communities. *Journal of Studies on Alcohol and Drugs, vol. 73, no. 6, pp. 968-975, 2012.*

Sunny, J.K., Lisa, A.M., Jeffrey. T. H. & Amarendra, K.D. (2017). Scaling Up Research on Drug Abuse and Addiction Through Social Media Big Data. *Journal of Medical Internet Research.*

Takeo, K., Tim, K.M., Mas & Raphael, C. (2015). Establishing a Link Between Prescription Drug Abuse and Illicit Online Pharmacies: Analysis of Twitter Data. *Journal of Medical Internet Research.*

Tao, Warren & Shimei. (2017). Social Media-based Substance Use Prediction. *arXiv:1705.05633*

Tim, K.M., Janani, K., Takeo, K. & Gert, L. (2017). Twitter-Based Detection of Illegal Online Sale of Prescription Opioid. *American Journal of Public Health*

Van Hoof, J.J., Bekkers, J. & van Vuuren, M. (2014). Son, you're smoking on facebook! College student disclosures on social networking sites as indicators of real-life risk behaviors. *Computers in human behavior, vol. 34, pp. 249-257, 2014.*

Wolpert, D.H. (1992). Stacked generalization. *Neural Networks, 5(2): 241-59.*

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

Xitong, Y. & Jiebo, L. (2017). Tracking illicit drug dealing and abuse on Instagram using multimodal analysis. *ACM Transactions on Intelligent Systems and Technology*.

Yiheng, Z., Numair, S. & Jiebo, L. (2016). Fine-grained Mining of Illicit Drug Use Patterns Using Social Multimedia Data from Instagram. *IEEE International Conference on Big Data.*

Yu, F., Moh, M., & Moh, T.S. (2016). Towards Extracting Drug-Effect Relation from Twitter: A Supervised Learning Approach. *International Conference on Big Data Security on Cloud (BigDataSecurity).*

Zahedi, F. & Zare- Mirakabad, M. R. (2014). Employing data mining to explore association rules in drug addicts. *Journal of AI and Data Mining.*

2012-2013 national survey on drug use and health: National maps of prevalence estimates, by state. *Technical report, National Survey on Drug Use and Health, 2014.*

KEY TERMS AND DEFINITIONS

Social Media: Websites and applications that enable users to create and share content or to participate in social networking.

Multimedia: Content that uses a combination of different content forms such as text, audio, images, animations, video and interactive content.

Data Mining: Process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics and database systems.

Demographic: Relating to the structure of populations.

Illicit Drug Use: Illegal usage of drugs that are illegal to make, sell or use.

Natural Language Processing: Automatic manipulation of natural language, like speech and text, by software.

Multimodal Analysis: Analysis of communication practices in terms of textual, aural, linguistic, spatial and visual resources.

In-degree: The number of arcs leading to this vertex.

Out-degree: The number of arcs coming from this vertex.

Neural Networks: Simple models of the way the nervous system operates, with neurons as basic units that typically organized into layers.

Classifier: An approach that classifies some meaningful thing.