Improving Prevention of drug addiction based on social media via clustering approach

Tagreed S. Alsulimani
Department of Management Information System
College of Business, University of Jeddah, Jeddah, Saudi Arabia
tsalsilimani@uj.edu.sa

Abstract
Drug addiction is a major problem in our societies, it affects any community and requires a serious solutions. The best approach to the problems of alcohol and drug is prevention. It is important to know that some interventions, including group therapy with other offenders, can do more harm than good in reinforcing consumption or antisocial behavior. Therefore, given the availability of data sources, an unsupervised study can be used to make decisions about the prevention and treatment of substance abuse. Hence, the role of social networks in the exploitation of biomedical knowledge, including clinical, medical and health informatics, the epidemiology of drug addiction, and the pharmacology of drugs, has become increasingly important in recent years. We also want to understand the relationship between a person's activities in social networks and behavior related to addiction.

Keywords: drug addiction, data analysis, social media, k-means

1. Introduction

Drug abuse is prevalent among young people and can be heavy consequences. The problem is treated well if it is taken early, but prevention is better than cure.
For example, It is estimated that one in 20 adults has consumed at least one in 2014. This represents 250 million people aged 15 to 64, approximately equivalent to the populations of Germany, France, Italy and United United Kingdom; that's a lot, but it does not seem have been increasing in the last four years proportionately to the world population [unseco 2016]. Based on the same source, in 2014, an estimated 183 million people would have used cannabis, which would always be the most commonly consumed globally, followed by amphetamines. In fact, more deaths, illnesses and disabilities are associated with drug abuse than any other preventable health condition. The major problem that the risk of drug use increases greatly during times of transition. For an adult, a divorce or loss of a job may increase the risk of drug use. Hence, Setting up of drug policies based on scientific data can, thanks to prevention and treatment measures example, mitigate the adverse consequences that consumption of drugs has for health. When based on scientific data, preventive measures, early intervention, treatment, care, recovery, rehabilitation and social integration and the entire care system for drug users reduce consumption and thus limit its impact on public health, which is one of the essential elements for the well-being of society.
In this paper, we report our effort on automated prevention of drug addiction based on social media. We apply an unsupervised study to automatically prevent any drug addiction.
The main contributions of this research include

1) Applying the state-of-the-art data mining approach to identify substance use- related social media documents. So far, there has not been much work on automated prevention of substance use- related social media.
2) Conducting comprehensive evaluations to demonstrate the effectiveness of the proposed method.

This paper is organised into five sections, this being the first one. Section 2 reviews the literature and presents the related work. Section 3 describes our synthetis result. We introduce our proposal on section 4. Section 5 starts with a description of the datasets and metrics used in experiments and followed by a deep investigation into the parameter tuning of the proposed method. Finally, Section 6 points out conclusions and future work.

2. State of the art : Drugs of abuse and social media

Social media users are increase every second and the number of active users suggested to be around 2.32 billion users in 2019 [1]. The huge distribution for different application of social media in smart phone and tablets enable everyone on the earth to chat and communicate with eachother's.

Many individuals start to post their daily activity through different social media application due to the wide use of social media among the world. Social media applications include: - WhatsApp, Snapshot, IMO, we chat, YouTube and others. This technology has a great impact on knowledge and education but there are also other dangerous effects. Several authors and researchers studied social media for medical, educational, statistical and other purposes.

The big problem with social media is that within very short time individuals can communicate with each other's and build a network to find illegal products such as weapons and drugs [2]. In addition, social networks can be used by terrorists to post some movies and photos for terrorist attack like ISIS and other terrorist groups. According to U.S. Central Command [3].

Threats include: - thieves, hackers, phishers/scammers, terrorists, intelligence spies and pedophiles. Thieve scan follow the internet users and determine their location and to know if someone is available at home to perform a crime.

Pedophilias using social media to trap young children for sex offenders (e.g MySpace). Phishers sending fake e-mail for internet users to update their information though using malicious software and collected user's sensitive data.
In addition,
Terrorists such as ISIS and Al Qaeda using internet to seek information about officers, authorized persons (work time, work location etc) [3].
Social media was used for medical studied for example it was used to study depression [4]. There is lot of arguments about advantages and disadvantages of social media A recent study suggested that reduce time for using social media to 10 minutes/day have significant effect in reduce mental illnesses such as depression, anxiety and fear [5]. Analysis of social media was an interested topic for many authors in fact, social network analysis (SNA) was started on 1991 when Malcom Sparrow developed social network analysis (SNA) as a tool for criminal intelligence [6]. Other studies on social media network analysis (SNA) focused on criminal's identification [7-8].
Social media application and technology can be use as a tool for drug traffic and smuggling through the Internet. However, many authors studied this phenomenon by using data mining to analyses social media for drugs of abuse traffics. A previous study focusses on the detection and analysis both illicit drugs (e.g. cocaine, heroin, marijuana etc) and prescription drug (e.g. Ritalin) abuse using tweets [9].Tweeter was used to determine the drugs abusers through their tweets [9]. Death related to opiate use & abuse increases every year in USA and it was suggested that there is more than 21% increase in death related to opiate and opiate like medication due to overdose [9] Tweeter was used also to determine abuse of opiates drugs such heroin, morphine and use of synthetic opiate prescribed drugs such as OxyContin, Ritalin, Vicodin. [10-11] The method for pervious study was depend on four steps

1- The tweets were collected by using Application Programming Interface (API) which allows applications to communicate with each other.

2- All data will be manged through database management systems (DBMSs). It provides users with a systematic way to update and manage data.

3- Tweets are reused and filter and inserted into data analysis and artificial intelligent systems to identify tweets contain drugs of abuse words

4- Systems will be used in different drug abuse (illegal & medical) monitoring services

In addition to tweeter other social media was studied by many authors. Instagram was used as a tool to identify drugs of abuser through examining users' posts [12]. A previous study applied on 58000 posts and 100 account for drug abusers [12]. Hashtags were examined by using NYSAGO from New York State Attorney General's office (include1000 drugs of abuse related posts). All Hashtags set-up-to- date by using Apriori algorithm [13].

Results showedthat percentage of cases consuming cannabis, pills and cough syrup were 72%, 14% and 13% respectively [12].

3. Proposed Approach: K-Means for drug addiction Prevention

Clustering has been one of the most widely studied topics in data mining and one of the popular unsupervised machine learning algorithms.
Clustering refers to techniques for grouping similar objects in clusters. Most of the initial clustering techniques were developed by statistics or pattern recognition communities, where the goal was to cluster a modest number of data instances.
Developing clustering algorithms to effectively and efficiently cluster rapidly growing datasets has been identified as an important challenge.
Clustering algorithms partition a set of objects, users in our case, into groups called clusters. A cluster refers to a collection of data points aggregated together because of certain similarities.
You'll define a target number k, which refers to the number of centroids you need in the dataset. A centroid is the imaginary or real location representing the center of the cluster.
The classical algorithm K-means was introduced and drawing by Hartigan [14]. This algorithm is a classification method that allows to reserve a set of data in k homogeneous classes ; k is a user fixed value. It affects each object, randomly, to a region and we iterate as follows: the centers of the different groups are recalculated and each object is assigned to a new group, based on the nearest center. Convergence is reached when the centers (also called centroids) are fixed.
K-Means algorithm has many advantages, including the implementation facility and the rapid convergence compared to other data mining approaches.

3.1 K-means Algorithm

K-means algorithm is an iterative algorithm that tries to partition the dataset into K predefined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the inter-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

The way k-means algorithm works is as follows:

1. Specify number of clusters K.
2. Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
3. Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.
   - **3.1.** Compute the sum of the squared distance between data points and all centroids.
   - **3.2.** Assign each data point to the closest cluster (centroid).
   - **3.3.** Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

The following figure (see Figure 2) presents an example of the K-means application on some data.

3.2 K-means for drug addiction prevention

Social media is the collective of online communications channels dedicated to community-based input, interaction, content-sharing and collaboration. Social media has become a central point of a person's daily life for many people around the world with the ability to be connected to these sites through access to cellphones, tablets, and computers [15]. There are various social networking sites available on internet like LinkedIn, Facebook, Instagram, Twitter, ... Hence, it is a tedious task to analyze the complex data. It is of great importance for academic and business to analyze such online social communities and predicting their behavior. The following figure (see Figure 1) presents the different steps of the proposed approach

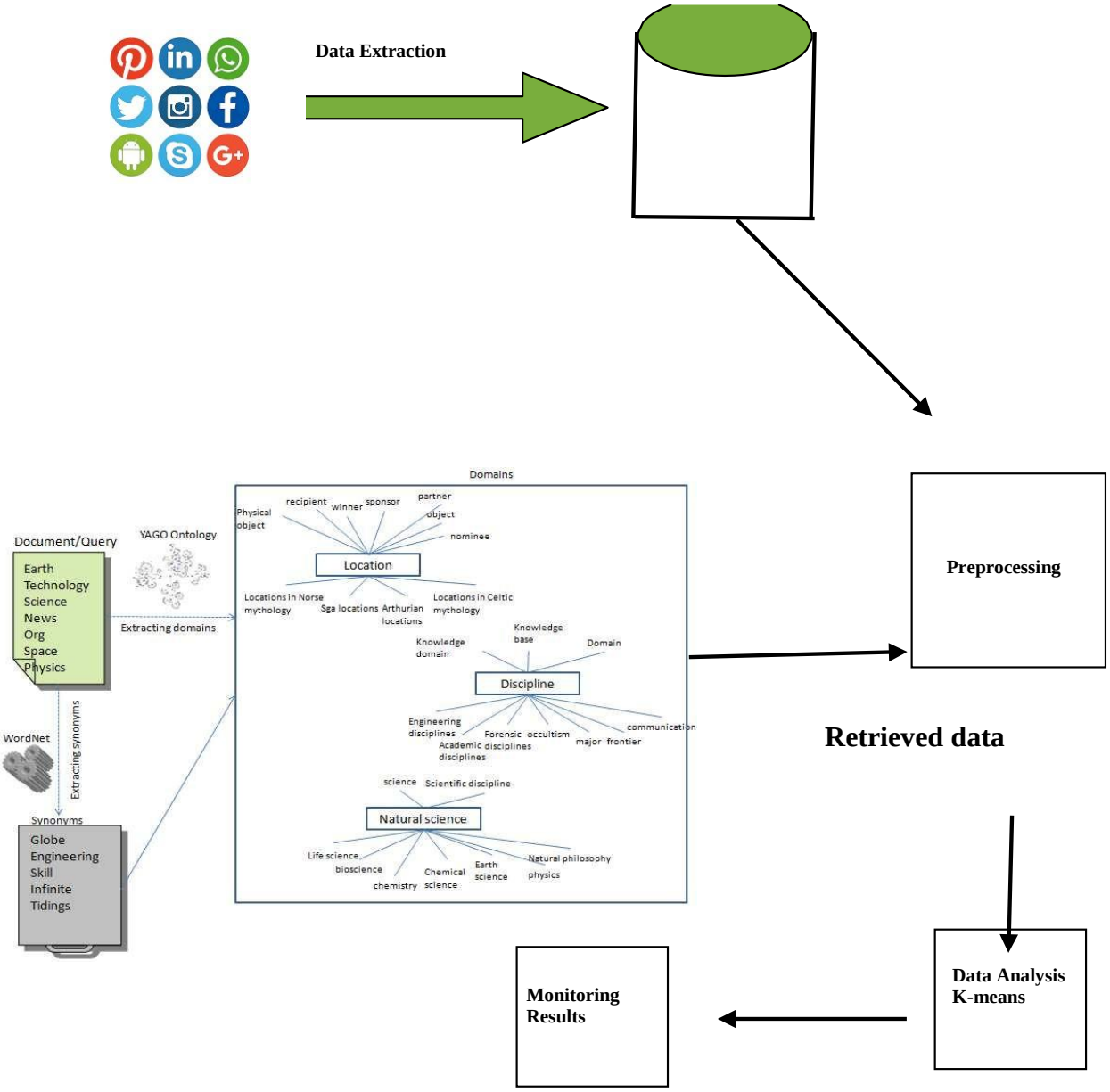

Figure 1: The proposed model

Data Retrieval from social media

Data retrieval involves information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, data mining techniques including the link and association analysis, visualization, and predictive analytics. We can use to retrieve any tool to retrieve data. Observe the comments variables that are twitted in the page for the opinion of the peoples. Now the data are moved to the next level data preprocessing to remove the repeated words and unwanted data's.

Data Preprocessing

Text preprocessing aims to make the input documents more consistent to facilitate text representation, which is necessary for most text analytics tasks. This is the most important part as social media comments do not have any specific format.
Stop word removal eliminates words using a stop word list in which the words are considered more general and meaningless. Stemming algorithms differ in respect to performance and accuracy and how certain stemming obstacles are overcome. The next step is to create corpus vector of all the words.
Once we have created the corpus vector of words, the next step is to create a document term matrix *DTF*.

Adapted K-means

The different steps of our appraoch is described below:

1. All documents in the initial collection called *DTF* is partitioned according to the number of documents processed at each iteration n.
2. Clustering of centroids: Centroids already generated previously, are grouped by K-Means algorithm to produce centroids cluster.
3. Mapping between document clustering and centroids clustering: This step is the intermediate step between the clustering of documents and the clustering of centroids to obtain the final clustering documents

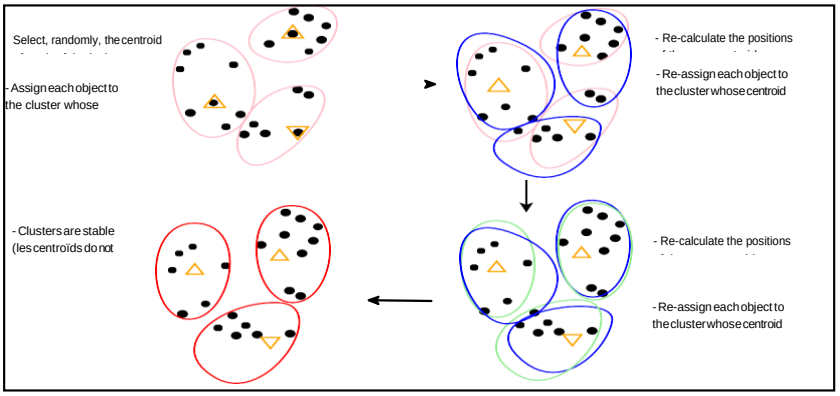

Figure 2: Adapted K-means

4. Conclusion

In this paper, we proposed a new algorithm for clustering social data. The algorithm is built based on a prepared data from social media to prevent . The basic idea relies on the automated prevention of drug addiction based on social media. We apply an unsupervised study to automatically prevent any drug addiction. An interesting research direction is therefore to do an experimental study on real data.

Acknowledgment: I would like to express my special thanks and gratitude to Doctor Khedija Belgacem Arour, who contributed in stimulating and suggestions for this paper.

References

Statista.com. https://www.statista.com/statistics/272014/global-social-networks- ranked-by-number- of-users/.

Drugabuse.com. https://drugabuse.com/featured/instagram-drug-dealers/.

US. central command. https://www.centcom.mil/VISITORS-AND- PERSONNEL/SOCIAL-MEDIA-SECUR ITY/.

Munmun C., Michael G., Scott C., &Eric H. Predicting Depression via Social Media. (2013) 2. Cody B., Jennifer G. This is your Twitter on drugs. Any questions? MSM'15, WWW (2015).

Melissa G. Hunt, Rachel Marx, Courtney Lipson, and Jordyn Young (2018). No More FOMO: Limiting Social Media Decreases Loneliness and Depression. Journal of Social and Clinical Psychology: Vol. 37, N o . 1 0 , p p .751-768. https://doi.org/10.1521/jscp.2018.37.10.751

Sparrow, Malcolm K.. "The application of network analysis to criminal intelligence: An assessment of the prospects." (1991).

Quiggins P. Police catch two fugitives with help of Facebook. http:// www.wkyt.com/home/headlines/Police_catch_two_fugitives_with_help_of_Fac eboo k_139556273.html. (2012).

Burcher, Morgan & Whelan, Chad. (2017). Social network analysis as a tool for criminal intelligence: Understanding its potential from the perspectives of intelligence analysts. Trends in Organized Crime. 21. 1-17. 10.1007/ s12117-017-9313-8. file:///C:/Users/1017724541/Desktop/Tagreed/Social %20network%20analysis%20as%20a%20tool%20for%20criminal %20intelligence.pdf.

Phan, Nhat Hai & Chun, Soon & Bhole, Manasi & Geller, James. (2017). Enabling Real-Time Drug Abuse Detection in Tweets. 10.1109/ICDE.2017.221.).

A. Sarker, K. O'Connor, R. Ginn, M. Scotch, K. Smith, D. Malone, G. Gonzalez, "Social media mining for toxicovigilance: automatic monitoring of prescription medication abuse from twitter," Drug Saf. 2016 Mar; 39(3):231-40. doi: 10.1007/ s40264-015-0379-4. PMID:

26748505.

L. Shutler, L. S. Nelson, I. Portelli, C. Blackford, J. Perrone, “Drug use in the Twittersphere: a qualitative contextual analysis of tweets about prescription drugs,” J Addict Dis. 2015;34(4):303–10.).

hou, Yiheng & Sani, Numair & Lee, Chia-Kuei & Luo, Jiebo. (2016). Understanding Illicit Drug Use Behaviors by Mining Social Media.).

Agrawal R. and Srikant R. Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases. (1994).

H. Bock, Origins and extensions of the k-means algorithm in cluster analysis, Electronic Journal for History of Probability and Statistics, 4 (2008), pp. 1–18.

V. Gurusamy et Al, Mining the Attitude of Social Network Users using K-means Clustering. International Journal of Advanced Research in Computer Science and Software Engineering, Volume 7, Issue 5, May 2017, pp. 226-230.