ISSN:2348-2079

Volume-5 Issue-1

International Journal of Intellectual Advancements and Research in Engineering Computations

Identifying alcohol and other drug usage among adolescents in India and suggesting a mechanism for rehabilitation using social media contents

1. Dr.C.S.Kanimozhiselvi, Associate Professor
2. S.P.Shangeetha, K.Shanmugapriya, G.Sivaram, UG Student
Department of computer science and Engineering
Kongu Engineering College, Perundurai
kshanmugapriya106@gmail.com

Abstract: Adolescent drug usage like tobacco, alcohol and other substances is a major public health problem that causes more annual deaths. Conventional methods for monitoring adolescent drug consumption are based on surveys, which have many limitations and are difficult to scale. With the growing popularity of social networking sites and the proliferation of mobile devices and camera phones, new opportunities and challenges emerge as people can now actively generate contents that provide a unique compilation of information that is more frequently updated and self-representative than traditional mode of data collection. Hence, a novel approach to monitoring underage drug use by analysing the content from social media networks in order to overcome many of the limitations of conventional approaches will be attempted in this proposal. This project aims to identify the conditions and behaviours of youth who use drugs from their social media activities and help them to come out of their problem by changing their mood with the appropriate content available from the social media. This project will also help the policy makers and health professionals to enter into the world of youth and addressing their problem and to guide them in an efficient way than the traditional means of counselling.

INTRODUCTION

Data Mining is an emerging field in computer science which deals with knowledge discovery from raw data. data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytic tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.Text mining is the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text , deriving patterns within the structured data, and finally evaluation and interpretation of the output.Text analysis involves information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, data mining techniques including link and association analysis, visualization, and predictive analytics. The overarching goal is, essentially, to turn text into data for analysis, via application of natural language processing (NLP) and analytical methods.

RELATED WORK

Alcohol is the drug of choice among youth. According to the National Survey on Drug Use and Health (NSDUH), a survey carried out in 2013 by the U.S. Department of Health and Human Services, an estimated 8.7 million underage persons (aged 12 to 20) were

current drinkers,including 5.4 million binge drinkers (consuming 4 or more drinks per occasion for women or 5 or more drinks per occasion for men at least once in the past month) and 1.4 million heavy drinkers.Surveys are the most widely used data collection method for underage drinking. The MTF and NSDUH are the Federal Government's largest and primary tools for tracking youth substance use.All these surveys provide useful data for the understanding of alcohol problems and the elaboration of specific prevention campaigns but still, they suffer limitations that are intrinsic to surveys like sampling error margins, high costs in terms of time and money, low scalability and delayed results generation since data collection.

Pang et al (2002) examined whether it is sufficient to treat sentiment classification simply as a special case of topic-based categorization or whether special sentiment-categorization methods need to be developed. This approach used three standard algorithms: Naive Bayes classification, maximum entropy classification, and support vector machines (SVMs) for sentiment classification. In topic-based classification, all three classifiers have been reported to achieve accuracies of 90% and above for particular categories. This shows that sentiment categorization is more difficult than topic classification.

Turney (2002) measured the co-occurrences between a word and a set of manually selected positive words (e.g., good, nice, excellent and so on) and negative words (e.g., bad, nasty, poor and so on) using pointwise mutual information to compute the sentiment of a word.

Liu et al (2004) proposed a method to summarize all the customer reviews of a product. This summarization task is different from traditional text summarization because the users are much interested in the specific features of the product that customers have opinions on and also whether the opinions are positive or negative. Hence, the approac
does not summarize the reviews by selecting or rewriting a subset of the original sentences from the reviews to capture their main points as in the classic text summarization. It focuses on mining opinion/product features that the reviewers have commented on. The drawback is that there is no group features according to the strength of the opinions that have been expressed on them.

Ran Pang et al.(2015) have employed the state of the art computer vision algorithms for face analytics to acquire information on age, gender and race of teenagers involved in alcohol use and built a comprehensive language model to capture drinking related activities from the tags associated with Twitter photos.

Rex Sahaya Raj et al(2013) analyzed the socio economic and relationship problem of alcoholics residing in Trichy Slums in Tamil Nadu. The researchers collected data from 40 respondents by using Simple Random Sampling and Self Prepared Questionnaire based on the study dimensions and they have used traditional methods for analyzing the data collected.

The **main contributions** of this work are several folds:

1. We identify a demographically matching social multimedia platform in Twitter to study theimportant social problem of underage drinking;
2. We employ the state of the art computer vision techniques to identify the targeted population among Twitter users;
3. We exploit natural language understanding based on an extensive dictionary to extract activity signals;
4. We combine robust and complementary signals from multimodalities to discover behavior patterns

at a large scale and a fine granularity not seen before on this subject; and

5. We develop a novel alternative approach to surveys using social multimedia and obtain promisingresults for a public health problem.

DATA SOURCE:

Several features of Twitter make it a good choice for monitoring adolescent alcohol use. Today there are more than 23 million Twitter users 2 among which around 20% are aged between 16 and 24.Twitter is a free social networking and microblogging service which enables to broadcast

short messages to your friends, family, co–workers or so called “followers” in real-time.Atimeline on Twitter is a collection of tweets in chronological order. The Public Timeline consists of every public tweet made. When you tweet, you create your own timeline that people will see when they visit your

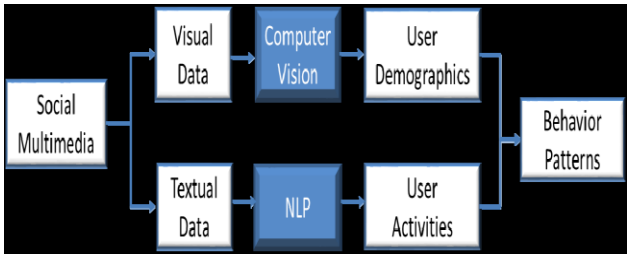
profile page. You can see your own timeline by clicking the Profile link in the top menu.ReTweeting is when someone repeats someone else's tweet, so their own followers can see the original message.

There are more than 284 million active Twitter users monthly, 80 percent of which are using mobile devices to tweet (Twitter). About 63 percent of Twitter users regard their smartphones as their primary tweeting device.Every second, on average, around 6,000 tweets are tweeted on Twitter (visualize them here), which corresponds to over 350,000 tweets sent per minute, 500 million tweets per day and around 200 billion tweets per year.Twitter allows you to send and read other users updates (known as tweets).Twitter messages(tweets) are limited to 140 characters (microblogging)

You can send and receive updates via the Twitter website, SMS(text messages), RSS (receive only), emails or a third party application.You can restrict delivery to your circle of friends (delivery to everyone is the default).You can use third party application such as Tweetie, Twitterrific, and Feedalizr to sendTwitter messages. You can search for people by name or user name, import friends from other networks, or invite friends via email.

METHODOLOGY

Twitter does not provide any personal information about its users (such as real name, age, gender, address).Therefore, most of our study will be based on extracting information out of the only data available: contents generated by users. Taking into account that Twitter is an text based social network where most of the users upload self-representative pictures and tweets that these are usually accompanied by descriptive hashtags, we can make use of both these elements to infer the missing data needed for our study. This *computational* data analytics framework is shown in Figure and generally applicable to problems that involve using social media to study user behaviors.



Text Analysis

The first, quickest and easiest way of filtering media and profiles is by analyzing the text attached to them. Pictures and videos contain descriptions and comments (which can also contain hashtags) and user profiles can also have some text such as a bio. We will use the bio and descriptions created by owners of accounts to detect the language spoken by the user with aid of existing Java libraries.We will also use customized dictionaries in order to filter specific contents requested to Twitter.set of keywords will be used for the *alcoholdictionary* (see appendix A). This dictionary will be used to

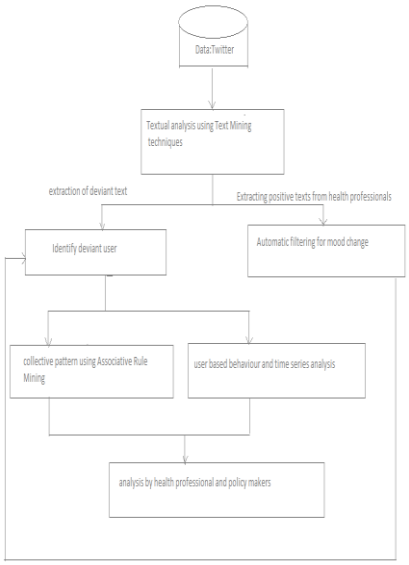
identify the person associated with alcohol consumption.If a user posts a picture with alcohol related tags such as "tequila" or "drunk", and considering most of the contents are usually self-referenced, it is reasonable to believe that the picture might have been taken during an alcohol consumption act where the user was involved. The procedure for creating the dictionary was iterative and manual. In the beginning, the dictionary only contained a few seed terms directly associated with drinking behaviours and conditions (drunk, drinking and alcohol).

PROPOSED FRAMEWORK:

Data from social media will be obtained from the API (Application Programming Interface) provided by social media sites. Social media contains different types of data – user information, connections between users, generated by users' content and etc. Each data type usually accessed by separate API.

The proposed system uses twitter reviews to extract aspect and mine whether the given opinion is positive or negative. Each review is split into individual sentences. A review sentence is given as input to data preprocessing. Data preprocessing consist of

stop word removal and Part Of Speech (POS) tagging.



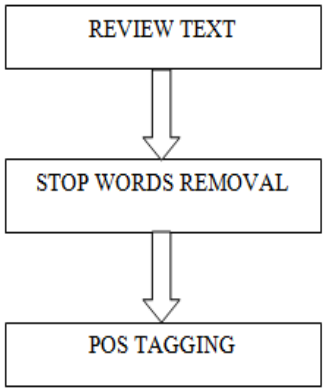
Stop word removal is used to remove irrelevant words. The POS tagging is used to generate the tag of each word. Next, it extracts aspect in each review sentence. And, the system extracts opinion words in customer reviews. Opinion can be classified in n-level orientation scale. Opinion orientation is an intended interpretation of the user satisfaction in terms of numerical values. It is used to identify whether it is positive or negative opinion sentence. It identifies the number of positive and negative opinions of each aspect. Also, it summarizes the importance of each aspect in reviews.

Data Preprocessing

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and lacking in certain behaviors or trends, and is likely to contain many errors. It is used to remove irrelevant information in reviews. Tasks in data preprocessing are

1. Data cleaning: fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.
2. Data integration: using multiple databases, data cubes, or files.
3. Data transformation: normalization and aggregation.
4. Data reduction: reducing the volume but producing the similar analytical results.
5. Data discretization: part of data reduction, replacing numerical with nominal ones.

The data preprocessing steps in the proposed work is shown in figure below



Stop Words Removal

Most frequently used words in English are not useful in text mining. Such words are called stop words.Stop words are language specific functional words which carry no information.It may be of types such as pronouns, prepositions, conjunctions. Stop word removal is used to remove unwanted words in each review sentence. Words like ‘is’, ‘are’,‘was’ etc. Stop words are removed to reduce indexing or data file size and improve efficiency. Reviews are stored in text file which is given as input to stop word removal. Stop words are collected and stored in a text file. Stop word is removed by checking review texts against stop words list.

POS Tagging

The Part Of Speech (POS) of a word is a linguistic category that is defined by its syntactic or morphological behavior. Common POS categories in English grammar are: noun, verb, adjective, adverb, pronoun, preposition, conjunction, and interjection. POS tagging is the task of labeling (or tagging) each word in a sentence with its appropriate part of speech. POS tagging is an important phase of opinion mining, it is necessary to determine the features and opinion words from the reviews.POS tagging can be done manually or with the help of POS tagger. Manual POS tagging of the reviews take lots of time. Here, POS tagger is used to tag all the words of reviews. Stanford POS tagger is used to tag each word in a review sentence. The noun word like ‘pictures’ are tagged as

'pictures_NNS'. Finally tagged sentence is stored in text file.

Aspect Extraction

All the features are extracted from the reviews and stored in a database then its corresponding opinion words are extracted from these reviews. An object is an entity which can be a product, service, person, event, organization, or topic. It is associated with a set of components or attributes, called aspects of the object. Each component may have its own set of aspects.

Apriori algorithm is used to find all frequent itemsets using minimum support count. Each sentence is assigned as single transaction. Noun Words in each sentence is assigned as item sets for single transactions.

Opinion Words Extraction

Opinion words are in the form of adjectives and adverbs. Example: Opinion words are 'good', 'bad', 'amazing', 'not good' etc. Opinion words are extracted after POS tagging and stored in separate text file.

Sentence and Aspect Orientation

Steps are as follows:

1. The positive and negative opinion words and review sentences are stored in text file.
2. Split the sentence into the combination of words. It means first combination of two words and then single words.
3. First compare the combination of two words, if matched then delete that combination from the opinion. Again start comparing for the single words.
4. Initially, the probabilities of all the labels are zero [positive=0, negative=0]. Based on opinion, the probabilities of positive and negative labels get incremented. After comparing all the words of the sentence, the found probabilities of the labels are compared in the following manners.

   i. If the probability of positive label is greater than the negative, then the sentence or opinion is positive.
   ii. If the probability of negative is greater than the positive, then the sentence or opinion is negative.
   iii. If the probability of positive minus probability of negative is zero, then it is neutral.

The identified deviant users information is forwarded to health professionals. To address the youth with deviant behavior, we can take measures to change their moods. In order to do that, the posts from psychologists, health professional and rehabilitated drug users will be extracted from social media. The post may be text/stories or videos relevant to the rehabilitation. The posts which will induce the positive thoughts and depict the impact of drug usage will be filtered and communicated to the users with deviant behaviours

CONCLUSION

Adolescence is the time during which people develop and form their crucial values, personality traits, and beliefs. Hence, as deviant behaviors occur during adolescence, it is important to guide adolescents away from such behaviors and back to normal behaviors. Moreover, although there are various kinds of deviant behavior, most of them would either directly or indirectly affect youths.

We hope our findings will prompt health authorities to come up with better prevention strategies, in tackling risky and addictive behaviours linked to drug usage in young people. We believe scanning messages and photos on social media could help identify those who are most vulnerable.

We attempt to give a solution that will help the policy makers and health professionals to enter into the world of youth and addressing their problem and to guide them in an efficient way than the traditional means of counselling.

This will help the adolescent drug user to understand their problems, the implications and benefits of overcoming drug addiction.

We believe that our findings will definitely aid policy-makers in developing interventions to target the most at-risk populations – particularly adolescent and underage students with strong tobacco and alcohol identities.

References

[1]. 1.https://news.developer.nvidia.com/researchers-are-using-gpus-to- monitor-underage-drinking-on-instagram/

[2]. Ran Pang, Agustin Baretto, Henry Kautz, and JieboLuo University of Rochester, Rochester "Monitoring Adolescent Alcohol Use via Multimodal Analysis in Social Multimedia",2015

[3]. Rex Sahaya Raj. M, Sam Deva Asir R.M and Tamilenthi.S, A Study on Alcoholism Among Youth Residing In Trichirappalli Slums, Tamil Nadu, India**.** African Journal of Science and Research , 2(2): 01-07, 2013

[4]. Lima, Ana CES, and Leandro N. de Castro. "Automatic sentiment analysis of Twitter messages."Computational Aspects of Social Networks (CASoN), 2012 Fourth International Conference on.IEEE, 2012.

[5]. Mukkamala, RaghavaRao, et al. "Detecting corporate social media crises on facebook using social set analysis." 2015 IEEE International Congress on Big Data.IEEE, 2015.

Yakushev, Andrei, and Sergey Mityagin. "Social networks mining for analysis and modeling drugs usage." Procedia Computer Science 29 (2014): 2462-247.