IJRSET MARCH 2019 Volume 6, Issue 3 Pages: 12-18

ISSN 2394-739X

International Journal for Research in Science Engineering and Technology

IDISORDER DETECTION USING MACHINE LEARNING

[1] Philomina Mary Jenisha, [2] Sruthi Jayaraman, [3] Yuvasri R, [4] D. Jayakumar, ME
[4.] Associate Professor,
[1, 2, 3, 4] Computer Science,
[1, 2, 3, 4] R.M.D Engineering College,
[1, 2, 3, 4] Thiruvallur, TamilNadu.

ABSTRACT: Increased prevalence of social media usage across various age groups and social sectors and boom in popularity of cross media has led to the various muddles such as social network mental disorders (SNMDs) like Cyber-Relationship Addiction, Information Overload, and Net Compulsion etc. Based on a person's usage pattern on social media, it is possible to predict of the person suffers from one of the 3 disorders mentioned above. In our project, we introduce a machine learning framework, namely Social Network Mental Disorder Detection (SNMDD) for mining online social behavior. It makes use of data sets extracted on browsing patterns to predict the presence of the 3 types of disorders mentioned above.

Keywords: Machine Learning, Datamining, Naïve Bayes, K-means, Cyber-Relation addiction, Net Compulsion, Information Overload.

1. INTRODUCTION

Increased prevalence of social media usage across various age groups and social sectors and boom in popularity of cross media has led to the various muddles such as social network mental disorders (SNMDs) like Cyber-Relationship Addiction, Information Overload, and Net Compulsion etc. Based on a person's usage pattern on social media, it is possible to predict of the person suffers from one of the 3 disorders mentioned above. In our project, we introduce a machine learning framework, namely **Social Network Mental Disorder Detection (SNMDD)** for mining online social behavior. It makes use of data sets extracted on browsing patterns to predict the presence of the 3 types of disorders mentioned above. Several employers in today's world are interested in performance profiling of prospective employees based on social media behavior. This will also heavily influence the hiring decision.

OBJECTIVE:

The objective of the project is to develop a machine learning based model which will use certain usage patterns as the independent variables to predict the 3 classes of SNMD (the dependent variable)

APPROACH:

Machine Learning, Support vector machine, Naïve Bayes method

KEY LEARNINGS:

The way of predicting the 3 types of disorders using Confusion matrix which is constructed based on false positive and false positive results.

PROBLEM STATEMENT & OBJECTIVE:

Previous work in psychology has identified several crucial mental factors related to SNMDs, they are mostly examined as standard diagnostic criteria in survey

questionnaires. Detection based on questionnaires could be biased and time consuming and could lead to incorrect and delayed detection of these disorders.

To automatically detect potential SNMD cases of OSN users, extracting these factors to assess user's online mental states is very challenging. For example, the extent of loneliness and the effect of disinhibition of OSN users are not easily observable. Therefore, there is a need to develop new approaches for detecting SNMD cases of OSN users, which could potentially be automated to provide on-the-fly predictions which are also reliable.

The objective of the project is to develop a machine learning based model which will use certain usage patterns as the independent variables to predict the 3 classes of SNMD (the dependent variable).

LITERATURE SURVEY:

Research on mental disorders in online social networks receives increasing attention recently. Among them, content-based textual features are extracted from user-generated as blog, social for sentiment analysis and topic detection. Chang et. al employ an NLP-based approach to collect and extract linguistic and content-based features from online social media to identify Borderline Personality Disorder and Bipolar Disorder patients extract the topical and linguistic features from online social media for privacy issue, different functions on different etc. We propose a novel tensor-based approach to address the issues of using heterogeneous data and incorporate domain knowledge in SNMD detection depression patients to analyze their patterns and analyze emotion and linguistic styles of social media data for Major Depressive Disorder (MDD). However, most previous research focuses on individual behaviors and their generated textual contents but do not carefully examine the structure of social networks and potential Psychological features. Moreover, the developed schemes are not designed to handle the sparse data from multiple OSNs. Our framework is built upon support vector machine, which has been widely used to analyze OSNs in many areas.

2. OVERVIEW OF SNMD

The three types of disorders are defined as follows:

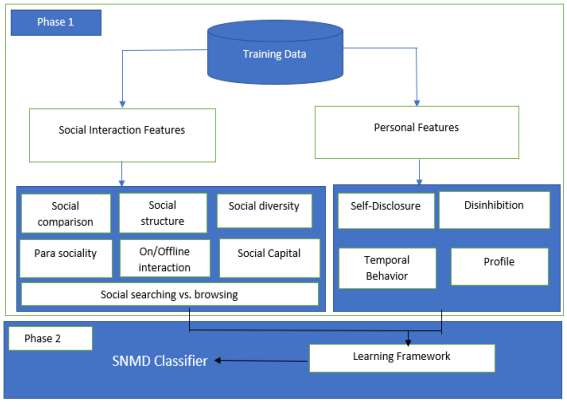
- **Cyber Relationship Addiction (CR),** which includes addiction towards cross media.
- **Net Compulsion (NC),** which includes compulsive online social gaming or gambling
- **Information Overload (IO),** which includes addictive surfing of user status and news feeds, leading to lower work productivity. We propose a two-phase framework, called Social Network Mental Disorder Detection (SNMDD).

SOCIAL STRUCTURE FEATURES:

In Sociology, each person in a social network belongs to one of the following three types of social roles: **influential users, structural holes, and normal users.** An influential user is the one with a huge degree. On the other hand, weaker connecting paths between groups are structure holes in OSNs, and researchers have demonstrated that structural holes usually have timely access to important information, e.g., trade trend, job opportunities, which usually lead to social success.

Therefore, the users with their roles as structural holes are more inclined to suffer from information overload for newsfeeds because they enjoy finding and sharing new and interesting information to various friends.

3. SYSTEM ARCHITECTURE



Social Interaction:

The way people talk and act with each other and various structures in society. It may include interactions such as a team, family or bureaucracy that is formed out of the need to create order within the interaction itself. It may also include interaction of social work. Social Interaction consists of many features, but regarding to this topic they are social comparison, social diversity, browsing, online/offline interaction.

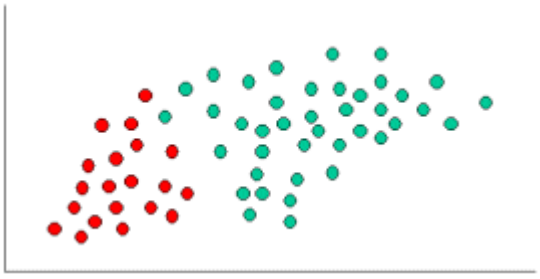
Personal features:

Personal data is any information that relates to an identified or identifiable living individual. Different pieces of information, which collected together can lead to the identification of a particular person, also constitute personal data. It includes features such as sincere, honest, loyal and understandable, etc.

4. MACHINE LEARNING TOOLS & TECHNIQUES

1. Naïve Bayes

The Naive Bayes Classifier technique is based on the so-called Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods.



To demonstrate the concept of Naïve Bayes Classification, consider the example displayed in the illustration above. As indicated, the objects can be classified as either GREEN or RED. Our task is to classify new cases as they arrive, i.e., decide to which class label they belong, based on the currently exiting objects. Since there are twice as many GREEN objects as RED, it is reasonable to believe that a new case (which hasn't been observed yet) is twice as likely to have membership GREEN rather than RED. In the Bayesian analysis, this belief is known as the prior probability. Prior probabilities are based on previous experience, in this case the percentage of GREEN and RED objects, and often used to predict outcomes before they actually happen. Thus, we can write:

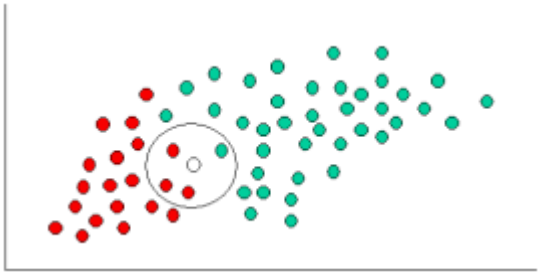
$$\text{Prior probability for GREEN} \propto \frac{\text{Number of GREEN objects}}{\text{Total number of objects}}$$

$$\text{Prior probability for RED} \propto \frac{\text{Number of RED objects}}{\text{Total number of objects}}$$

Since there is a total of 60 objects, 40 of which are GREEN and 20 RED, our prior probabilities for class membership are:

$$\text{Prior probability for GREEN} \propto \frac{40}{60}$$

$$\text{Prior probability for RED} \propto \frac{20}{60}$$



Having formulated our prior probability, we are now ready to classify a new object (WHITE circle). Since the objects are well clustered, it is reasonable to assume that the more GREEN (or RED) objects in the vicinity of X, the more likely that the new cases belong to that particular color. To measure this likelihood, we draw a circle around X which encompasses a number (to be chosen a priori) of points irrespective of their class labels. Then we calculate the number of points in the circle belonging to each class label. From this we calculate the likelihood:

$$\text{Likelihood of X given GREEN} \propto \frac{\text{Number of GREEN in the vicinity of X}}{\text{Total number of GREEN cases}}$$

$$\text{Likelihood of X given RED} \propto \frac{\text{Number of RED in the vicinity of X}}{\text{Total number of RED cases}}$$

From the illustration above, it is clear that Likelihood of X given GREEN is smaller than Likelihood of X given RED, since the circle encompasses 1 GREEN object and 3 RED ones. Thus:

$$Probability\ of\ X\ given\ GREEN \propto \frac{1}{40}$$

$$Probability\ of\ X\ given\ RED \propto \frac{3}{20}$$

Although the prior probabilities indicate that X may belong to GREEN (given that there are twice as many GREEN compared to RED) the likelihood indicates otherwise; that the class membership of X is RED (given that there are more RED objects in the vicinity of X than GREEN). In the Bayesian analysis, the final classification is produced by combining both sources of information, i.e., the prior and the likelihood, to form a posterior probability using the so-called Bayes' rule (named after Rev. Thomas Bayes 1702-1761).

$$Posterior\ probability\ of\ X\ being\ GREEN \propto$$

$$Prior\ probability\ of\ GREEN \times Likelihood\ of\ X\ given\ GREEN$$

$$= \frac{4}{6} \times \frac{1}{40} = \frac{1}{60}$$

$$Posterior\ probability\ of\ X\ being\ RED \propto$$

$$Prior\ probability\ of\ RED \times Likelihood\ of\ X\ given\ RED$$

$$= \frac{2}{6} \times \frac{3}{20} = \frac{1}{20}$$

Finally, we classify X as RED since its class membership achieves the largest posterior probability. The above probabilities are not normalized. However, this does not affect the classification outcome since their normalizing constants are the same. This is a well-established Bayesian method primarily formulated for performing classification tasks. Given its simplicity, i.e., the assumption that the independent variables are statistically independent, Naive Bayes models are effective classification tools that are easy to use and interpret. Naive Bayes is particularly appropriate when the dimensionality of the independent space (i.e., number of input variables) is high (a problem known as the curse of dimensionality). For the reasons given above, Naive Bayes can often out-perform other more sophisticated classification methods.

5. Model performance measures

The various models built, must be evaluated based on certain model performance measures to identify the most robust models. The choice of the right model performance measures is highly critical. The following model performance measures have been used to evaluate the models, based on the confusion matrix built for the predictions on the training and test datasets:

		Predicted	
Observed		**Negative**	**Positive**
	Negative (paid flag=0)	True Negative (TN)	False positive (FP)
	Positive (paid flag=1)	False negative (FN)	True positive (TP)

A. Accuracy:

Accuracy is the number of correct predictions made by the model by the total number of records. The best accuracy is 100% indicating that all the predictions are correct.

B. Sensitivity or recall:

Sensitivity (Recall or True positive rate) is calculated as the number of correct positive predictions divided by the total number of positives. It is also called recall (REC) or true positive rate (TPR).

$$Sensitivity = \frac{TP}{TP + FN}$$

For our dataset, it gives the ratio of actual paid customers by the total number of customers predicted as customers who will pay. Since our focus is on identifying customers who have a higher propensity to convert to paid subscription, it is important that sensitivity is high for the model.

C. Specificity:

Specificity (true negative rate) is calculated as the number of correct negative predictions divided by the total number of negatives.

$$Specificity = \frac{TN}{TN + FP}$$

For our dataset, specificity gives the ratio of actual free customers by the number of customers who are predicted as those who will not pay. If a model gives low specificity, it means that the number of correct prediction of free customers is relatively less. The impact of this is less compared to having a lower sensitivity. Hence, we can compromise on specificity to some extent.

D. Precision:

Precision (Positive predictive value) is calculated as the number of correct positive predictions divided by the total number of positive predictions.

$$Precision = \frac{TP}{TP + FP}$$

Precision tells us, what proportion of customers predicted as paid customers actually are paid customers. If precision is low, it implies that the model has lot of false positives. That is, many free customers have been predicted as customers who will pay. This is associated with an increasing calling cost by the tele-calling team who call high propensity customers to enable conversion.

E. Precision vs recall:

Recall or sensitivity gives us information about a model's performance on false negatives (incorrect prediction of customers who will pay as free), while precision gives us information of the model's performance of false positives (incorrect prediction of free customers as customers who will pay).
The optimum model for our problem statement is the one which minimizes false negatives without missing a lot on false positives. We will use the cost associated with false positive and false negative predictions to further evaluate the model.

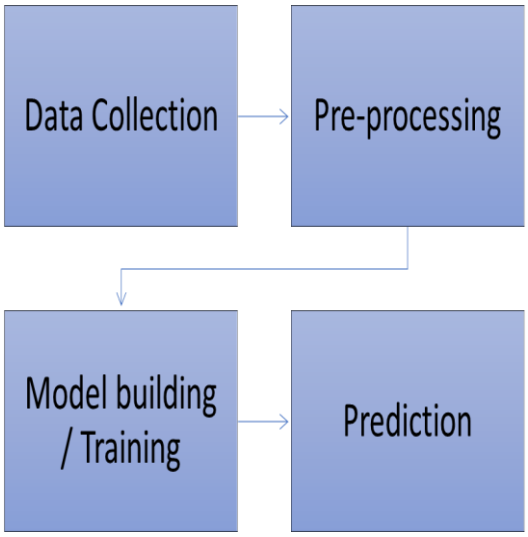
6. ABOUT THE DATASET

The dataset that we have used for this model consists of 10 attributes of social media users used as X variables and a label with 3 classes as the Y variable.

A. Data dictionary:

PROFILE STATUS	DESCRIPTION
AIN	How many times profile logs-in
AOUT	How many times profile logs-out
ONLINE	How much time profile is in online
OFFLINE	How much time profile is in offline
NSTRONG	How strong the network is available
NWEAK	How weak the network is present
EMOTICONS	How many emoticons are used
SELFIES	How many selfies are taken
DURATION	How much time profile is active in media
LABEL	Categorizing profile as CR, NC, IO.

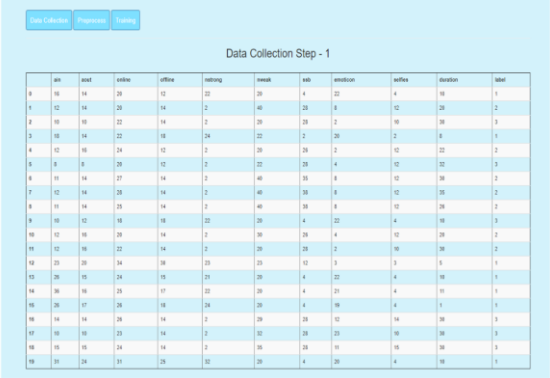
7. HIGH LEVEL APPROACH TO MODEL BUILDING:



8. DATA COLLECTION STAGE

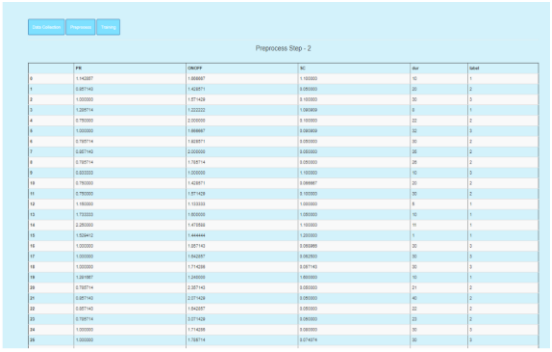
First stage of the process is data collection stage manually.
In data collection stage we collect data about various users from Facebook.
The collected data consists of the social interaction features and personal features, collected data are preprocessed that includes

finding null values and filling in missing values.



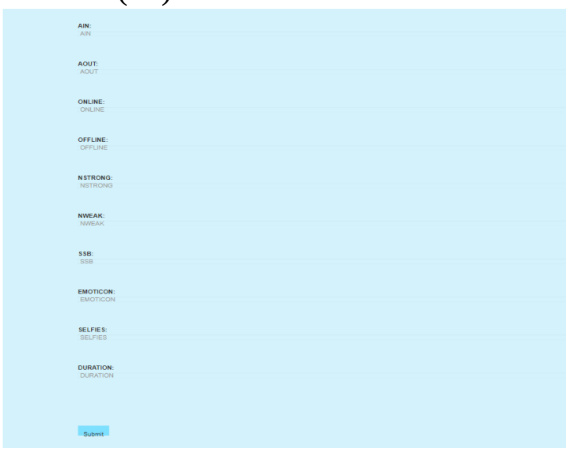
1. PREPROCESSING STAGE

In this stage, the data is been collected and classified according to kind of disorders. It includes emoticons. According to the symptoms and pre-diagnosed data. Admin relates them with each disorder.



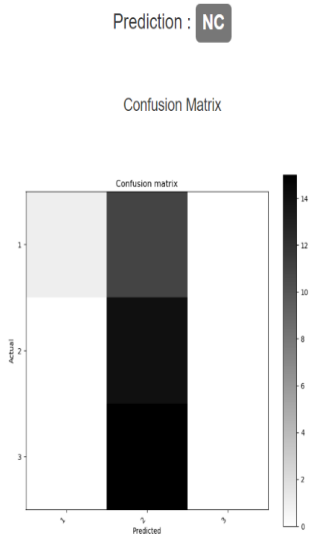
2. TRAINING STAGE

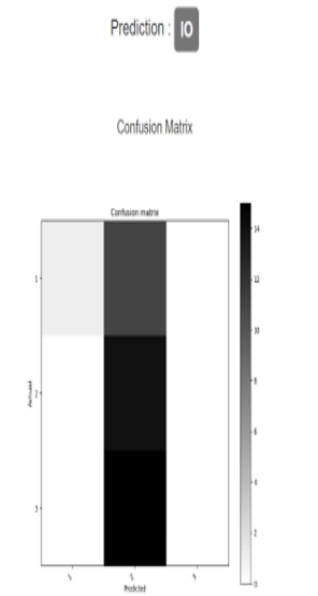
Once personal features are extracted next stage is training stage. In classification stage we use the Naïve Bayes for our classification algorithm. With the help of Naïve Bayes, classified user with mental disorder into one of following as: 1.Cyber Relationship Addiction (CR) 2. Net Compulsion (NC) 3.Information Overload(IO).



3. PREDICTION STAGE

Relapse is the state that a person is inclined to quickly revert back to the excessive usage of social media after an abstinence period, while tolerance is the state that the time spent by a person with SNMDs tends to increase due to the mood modification effect. It is worth noting that the above two mental states have been exploited to evaluate clinical addictions. We aim to use them to distinguish heavy users and addictive users because heavy users do not suffer from relapse and tolerance in use of OSNs. An app has been developed in python to get input variables from the user and predict the class of the Y variable (type of disorder). The code also generates a confusion matrix to show the accuracy of the prediction.





ACTIONABLE INSIGHTS

When predicting in the social media it is useful various fields of workers like

1. The employers can identify whether the processed employees have any social media related disorders.
2. The patients who are undertaking psychiatrist can use it as additional tool for diagnosis.
3. Can be overall used by person for personality profile.

CONCLUSION

In this paper, we make an attempt to automatically identify potential online users with SNMDs. We propose an SNMDD framework that explores various features from data logs of OSNs and a new tensor technique for deriving latent features from multiple OSNs for SNMD detection. This work represents a collaborative effort between computer scientists and mental healthcare researchers to address emerging issues in SNMDs. As for the next step, we plan to study the features extracted from multimedia contents by techniques on NLP and computer vision and apply it for private accounts, which can be incorporated in Instagram and Twitter.

REFERENCES

[1]. G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. (references)

[2]. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3]. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4]. K. Elissa, “Title of paper if known,” unpublished.

[5]. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[6]. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7]. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.