A Survey on Mining Techniques For Depression Prediction

Sneha Agrawal[1], Prof. R. J. Bhardwaj[2], Dr. S.B. Chaudhari[3]
[1, 2, 3] Dept of Computer Engineering
[1, 2, 3] Trinity College of Engineering and Research, Pune

Abstract- *Mental health plays a vital role in our life and is a crucial part in the concept of health. Depression cures earlier when detected at initial stage. Nowadays modern methods are involved to detect the state of the mood and the hidden activity of depression. There are number of elements and symptoms of depression. A study has been carried out for features selection methods. Depression can be detected with numerous methods but the results vary person to person and the methods being used for it. The conclusion that inferred was that the traditional method of questionnaire survey takes time to detect the depression. To extract and classify the features of depressants, mining techniques are studied. This paper examines and studies the literature about the mining methodologies, the depression and mental health to find out the relation between them.*

Keywords- Artificial neural network, data mining, depression, knowledge discovery, term frequency, text mining

I. INTRODUCTION

A significant component of human health is mental health. To discern mental state of a human depends on how he responds to the questionnaires, the accuracy of the answers given by him, the ability of memory, and the originality of the answers [1]. According to the definition of World Health Organization: Mental health is ”a state of well-being in which a person realizes his or her own capabilities, can handle the normal stresses of life, can work profitably and fruitfully, and is able to contribute in a effective manner to his community[1]. Depression can also be expressed as a level of cognitive behaviour or emotional healthiness or an absence of mental disorder. Depression and anxiety are some common mental health problems where as similar to them which are uncommon include schizophrenia and bipolar disorder [1]. To determine the symptoms of the depression, the depressants must answer all the questions precisely in self-report questionnaires. But this is very time consuming [2].” Today with the increasing challenges, daily stresses and pressure, more and more people face mental illness, and need personal attention and care. The mental illness can be determined so easily. However, if detected earlier the treatment can work well. It is essential therefore to screen the depression as soon as possible [2]. Depression has known to be contributor of unhealthy state of mind, loss of life (mostly by attempting to suicid) and financial loss and destruction of property [2].

According to the survey conducted, mining and soft computing can solve the problems of traditional methods. These techniques are broadly used resulting in fraud detection [2], handling of misconceptual data, scientific discovery [2] as well as medical domains. A fusion of mining methodologies, that is, text mining and data mining can extract unique features and detect depression with great accuracy. There are number of techniques of text mining and data mining as described in following sections.

II. REVIEW OF LITERATURE

Various web behaviours related to mental illness were analysed and the possibility of depressants using the social media or social web were discussed to diagnose the mental disorder under different conditions. This work was carried out by Fan Zhang, Tingshao Zhu, Ang Li, Yilin Li and Xinguo Xu in 2011.

In the year 2014, Hanen Mhamdi and Faouzi Mhamdi provided an overview for selection of methods and mainly focussed on Filter, Wrapper and the hybrid methods.

The paper of M. Janaki Meena and K R Chandran (2009) was reviewed for the study of text classification and machine learning.

Qi Luo presented the process for Knowledge discovery in Database (KDD) and proposed the data mining theory, data mining tasks, data mining technologies and data mining challenges. (2008)

The paper presented by Mrs. Sayantani Ghosh, Mr. Sudipta Roy, Prof. Samir K. Bandyopadhyay briefly reviewed the text mining algorithms, viz, Classification Algorithm, Association Algorithm, Clustering Algorithm stating their merits and demerits in the year (2012)

A study of the model developed by Putthiporn Thanathamathee was carried out. The study inferred that screening depression and predicting depression is based on the technique of feature selection.

To study the text mining approach, ie TF-IDF, LSI, Multiword in Information Retrieval and Text Categorization, a paper of Wen Zhang, Taketoshi Yoshida, Xijin Tang was reviewed.

III. PROPOSED WORK

1) Prediction of depression along with its level of severity
2) Combination of text mining and data mining for extracting more accurate features.

3.1 DATA COLLECTION

A survey form of nearly 25-30 questions is being developed and will be circulated among 800-1000 teenagers for their feedback. This data will be the input to the system. And a scale will be predefined with the reference of Center of Epidemiologic Studies Depression Scale (CES-D)[2], Kutcher Adolescents Depression Scale – 16 items (KADS-16) [10] and Hamilton scale[9]. Hence, this will state the risk of depression in a person. A wordlist of depression and non-depression will be formed to determine the sentiments regarding depression.

3.2 METHODS AND TECHNIQUES

In this paper a model is proposed that uses depressants health record for extracting and selecting the features to predict the depression along with its severity level. The proposed procedure mainly relies on two prime steps. The first step is to identify the features that will help for the depression predictions using text mining. The second step is to predict the severity level of depression using data mining. The overall system design is therefore depicted in the following diagram:

The first step, feature extraction techniques considered is TF (Term Frequency). TF is a text mining approach. Initially, the users' comments are processed. The text mining process, therefore comprises of parsing, tokenizing the document, filtering the common stopwords. Then applying term frequency (TF) to it. The formula of term frequency is stated as –

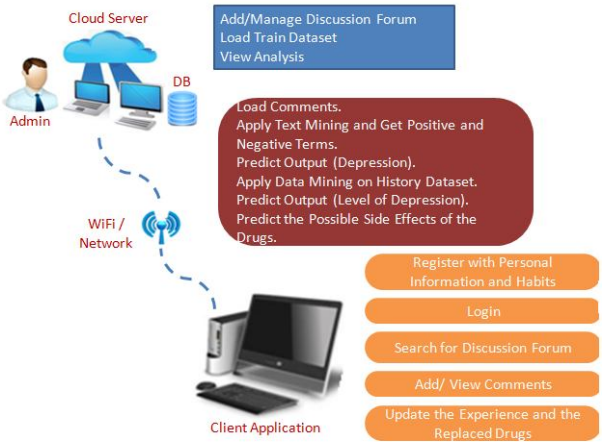

Fig 1: Proposed System Design

$$\text{Term Frequency (t)} = \frac{\text{Total number of term t in a document}}{\text{Total number of terms in a document}}$$

For more exact sentiment prediction the negation algorithm is applied and the depression is identified.

The pattern selection technique can be categorized into three groups: Filters, Wrappers and Hybrid Methods [3].

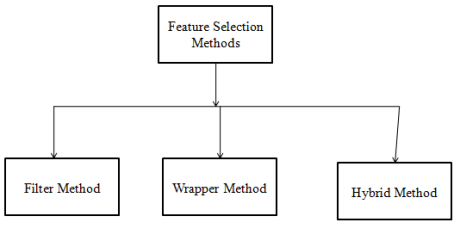

Fig 2: Classification of Feature Selection Methods

3.2.1 Filter Methods

It is a preprocessing step where the general characteristics are used to choose features without the induction algorithm. It is quick than the wrapper method and because of its independency of induction algorithm, it results in a better generalization [3] and is easy to interpret.

3.2.2 Wrapper Methods

These methods are too expensive for huge databases and generate a better performance than the filter methods because the process is improved for the classification algorithm used [3]. The wrapper method is classified as:

1) Sequential Algorithm (Forward and Backward): The process of Forward Sequential algorithm is: It starts with the empty set and adds one feature to continue with first step. This

gives the highest value for objective function. The next step onwards all the remaining features is individually added and a new subset is evaluated. This process is repeated until sufficient number of features is not added to the current subset. The process of Backward Sequential algorithm is: This process is similar to Forward Sequential algorithm, the only difference is that it start from complete set of the features and removes a feature at one time [3].

2) Heuristic Search Algorithm: This algorithm evaluates and different subsets are generated by searching in a search space and provides solutions to advanced problems.[3]

3.2.3 Hybrid Methods

This is the combination of the filter methods and wrappers methods. It gains the powerful features of both the approaches. It forms a better alternative when it comes for the better quality of space and time [3].

A. Text Mining

Text mining is the method that draws on information retrieval field, machine learning and computational linguistics [4]. It can also be termed as Text Data Mining, is a method of analyzing the text and a process to infer useful and relevant information from the unstructured text. Useful and relevant refers to combination of the relevance, novelty.

Text Mining Tasks and Stages

1) Text Categorization: With predefined categories, the documents are allocated to the dataset.
2) Text Clustering: Similar documents are grouped together and binded into groups knows as clusters.
3) Concept Mining: Discovers the concept from the original data, combination of text categorization and text clustering.
4) Information Retrieval: The information is retrieved that is pertinent to users' query.
5) Information Extraction: The useful information is extracted from large database.

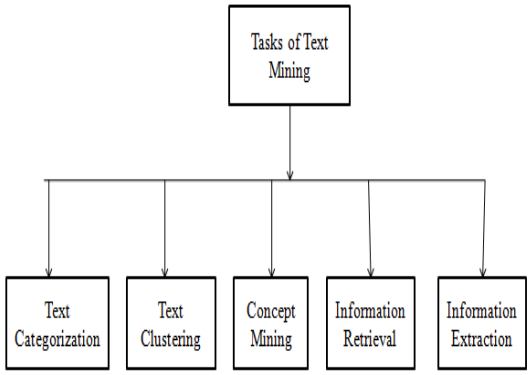

Fig 3: Classification of Text Mining Tasks

Text Mining has four stages: Information retrieval, Natural Language Processing, Data Mining and Information Extraction [4].

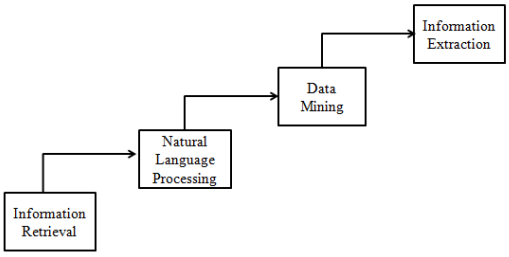

Fig 4: Stages of Text Mining

1) Information Retrieval: It is used in the libraries, where not the documents are required but only the digital required records are to be considered.
2) Natural Language Processing (NLP): It provides linguistic data that is required for the task of text mining. It is oldest method of Artificial Intelligence.
3) Data Mining: It is a phase of knowledge discovery and draws previously unknown information from the data.
4) Information Extraction: The useful information is extracted from large database. The data obtained is in the structured format.

The second step, the predicted sentiments and the output of previous step are considered by Artificial Neural Network (ANN), a data mining method. The ANNs are capable of learning, which takes place by altering weight values. Therefore, depending upon the output value of the algorithm value, the depression level can be identified.

B. Data Mining

Data can be anything like any fact, numbers, media, text, metadata which can be processed to deduce the design pattern. In this world of internet the huge amount of data is day by day increasing. It can be present in any format and numerous databases. The data can be operational (transactional), nonoperational, meta data [5].

Data Mining Tasks and Technology

The data mining tasks differ from the text mining because of presence of numerous patterns in vast databases. Distinct methods and techniques are required to search distinct kinds of knowledge patterns. Thus, it can be grouped as [5]:

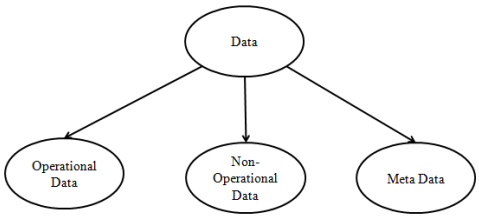

Fig 5: Different Types of Data

1) Summarization: To aggregate the data, it is summarized, that is an abstract is generated. It provides an overview.
2) Classification: It is constructed by examining the relationship between the variables and the classes.
3) Clustering: Similar documents are grouped together and binded into groups knows as clusters. The derived clusters are generally unknown and hence are uncovered.
4) Trend Analysis: The time series data is generated. Example: Share market prices.

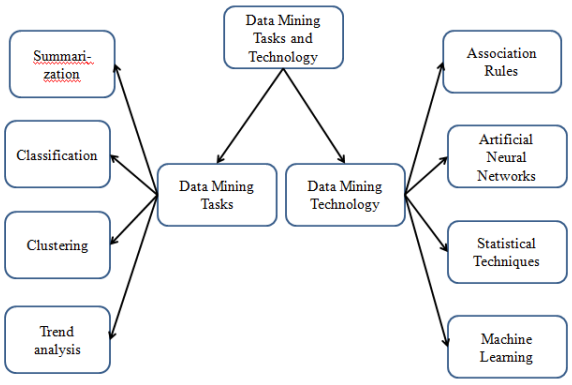

Fig 6: Data Mining Tasks and Technology

The data mining technology is classified as: Association Rules, Artificial Neural Networks, Statistical Techniques and Machine Learning [5].

1) Association Rules: To disclose the nature and the relation of the frequency, associate rule generator generates corresponding association rules. It also reveals the relationship between the data entities.
2) Artificial Neural Network: It used for parallel computing and capability of good generation of data, a learning framework is required.
3) Statistical Techniques: Statistical summarization, linear regression methods are included.
4) Machine Learning: It allows the generation of if-then rules for the problems of classification (crisp and fuzzy decision), regression, and temporal trees. A core part of data mining is turning the raw non-useful data into knowledge. Data mining is the analytical tool which analyzes the data from distinct dimensions, summarizing the relationship identification. It is a process of finding the relation amongst the data.

IV. CONCLUSION

Depression has now become a very serious matter and awareness has to be spread among the people. It is important to treat depression at its early stage because it becomes severe slowly. To diagnose depression is a difficult task and therefore an application to detect and predict depression is required. Therefore various papers have been reviewed to find out the techniques to develop such softwares. The distinct methods are studied. The feature extraction methods, text processing methods and data mining methods are studied into detail.

V. ACKNOWLEDGMENT

I would like to express my sincere gratitude to all the people who have guided, supported and helped me directly or indirectly; without whom this work would have not been possible. I am highly thankful to Prof. R J Bhardwaj for her guidance and consistent supervision. I would like to extend my special appreciation to my parents and all the staff members of Trinity College of Engineering and Research for their kind co-operation and encouragement which help me in completion of this work.

REFERENCES

[1] Fan Zhang, Tingshao Zhu, Ang Li, Yilin Li, Xinguo Xu, ”A Survey of Web Behaviour and Mental Health”, IEEE, 2011
[2] Putthiporn Thanathamathee,”Boosting with Feature Selection Technique for Screening and Predicting Adolescents Depression”, IEEE, 2014

[3] Hanen Mhamdi, Faouzi Mhamdi,"Feature Selection Methods for Biological Knowledge Discovery: A Survey", 25th International Workshop on Database and Expert Systems Applications,IEEE, 2014
[4] Mrs. Sayantani Ghosh, Mr. Sudipta Roy, Prof. Samir K Bandyopadhyay," A Tutorial review on Text Mining Algorithms", International Journal of Advanced Research in Computer and Communication Engineering, Vol 1, Issue 4, June 2012
[5] Qi Luo,"Advancing Knowledge Discovery and Data Mining", 2008 Workshop on Knowledge Discovery and Data Mining, IEEE, 2008
[6] M Janaki Meena, K R Chandran, "Naive Bayes Text Classification with Positive Features Selected by Statiscal Method", ICAC, IEEE, 2009
[7] Padhraic Smyth,"An Information Theoretic Approach to Rule Induction from Databases", IEEE Transaction on Knowledge and Data Engineering, Vol 4, No. 4, August 1992
[8] Wen Zhang, Taketoshi Yoshida, Xijin Tang, "TFIDF, LSI and Multi-word
[9] http://www.psychcongress.com/saundras-corner/scales-screeners/depression/hamilton-depression-rating-scale-ham-d
[10] http://www.shared are.ca/files/Kutcher_depression_scale_KADS11.pdf

1. Miss. Sneha Agrawal
She has enrolled herself for the degree of Masters of Engineering in the Department of Computer Engineering, Trinity College of Engineering and Research, Pune, India.

2. Prof. R. J. Bhardwaj
She is the Professor of Department of Computer Engineering, Trinity College of Engineering and Research, Savitribai Phule Pune University, Pune, India

3. Prof. S. B. Chaudhari
He is the Head of the Department of Computer Engineering of Trinity College of Engineering and Research, Savitribai Phule Pune University, Pune, India