International Journal of Academic Research and Development
ISSN: 2455-4197
Impact Factor: RJIF 5.22
www.academicsjournal.com
Volume 3; Issue 2; March 2018; Page No. 933-938

Intelligent data monitoring and mining in online social depression related networks

S Rashida[1], S Khader Basha[2]
[1] PG Scholar, Department of C.S.E, SSITS, JNTUA, Andhra Pradesh, India
[2] Assistant Professor, Department of C.S.E, SSITS, JNTUA, Andhra Pradesh, India

Abstract
As per health care perception Depression is a wellbeing concern at global range. In today social media's enhancement allows the people those who affected can share their experiences through posts. Such experiences are stored in database and can be extract and analyze to assist the precautions for others or recall the drugs from side effects, and other service improvements in their particular disease treatment. In those aspects social websites related to depression are helpful to extract knowledge or monitor on various types of drugs and its side effects and also for sharing their experiences on depression. We have taken a weighted edge network model for representing social networks activities. The proposed work undergoes with the three steps. The first step is for user activity monitoring and followed by network clustering and module analysis. Whoever the person like a specific posts belongs to a group and those who are not are belong to other group. We implemented the stop word technique here which is helpful in avoiding the misleading communication on the posts and efficient interaction of user. The statistical analyses of such user interactions are beneficial for health networks to acquire more knowledge on particular disease. This approach enables us all the gatherings took a part and for healthcare improvements in future to the patients of that disease.

Keywords: data mining, online fora, depression, stop-words technique

1. Introduction
Depression is a disease and said as a major contributor to the world wide suicides that happened especially in middle to low income countries. For instance, like India based on the study of World Health Organization (WHO) [1]. As per WHO basis of 2015 in its recent global health estimation about depression in 2015 stated that about 5 crore Indians are depression suffering one's, whereas three crore people are suffered from anxiety disorders. The report is entitled as "Depression and Other Common Mental Disorders-Global Health Estimates" said about two-thirds of global suicides are happening especially in low and middle income countries like India in 2015. The document of WHO shows that 322 million people are with depression and approximately half members from them live in South East Asian and Western Pacific region, reflecting relatively large populations of India and China [3]. Total people with depression in world are 322 million. Among them there estimated people with depression increased by 18.4% between 2005 and 2015. As per WHO figures in 2015 the number of depressive disorder cases are 3 crores of population whereas 3 crore of population is with anxiety disorders. It also said that "Suicide occurs throughout the lifespan and was the second leading cause of death among 15-29 years old globally in 2015". Depression is the main factor for causing disability in worldwide and is wellbeing concern to the overall global burden of disease, WHO said and asserted women are affected by depression than men and is leads to suicides [4, 5].
Forums and social media websites are for depression by sharing the experiences of healthcare workers and patients to manage in their routine lives and responds to antidepressants. Such huge data offers greater chances for patients, healthcare organizations, and industry to enhance solutions through intelligent data mining, extraction and analysis [2]. A virtual social media networking environment consists of nodes and edges. Its contents are modeled and extracted with the help of computational tools are trendy which formulate expectations and buildup the user relationships. The information will be represented visually by graphical representation. A social network's structure is represented by a socio matrix. Topological parameters like node degrees and network densities elucidate particular dynamics inside a network and a particular algorithm tends to underlying information oriented structures. Finding those clusters enables node (or cluster) centered data mining. Such important data helps the community for improving services that relies on feedback from "smart" data mining of health related social media sites. There are various methods in literature that helps in data gathering from social media networks are lexion-based, supervised classification, and concept extraction [6]. The rest of methods use graph-based analysis [7], text-based analysis derived from a medical corpus [8], and a topic-model statistical analysis [9]. *Zhao et al.* [10] used text-based analysis (posts length, certain words frequency) and sentiment analysis to identify influential user's online cancer survival communities. In contrast our approach integrates weighted network models (for user activity representation), module(describing user interaction) and topological (user activity) analysis with sentiment and text analysis to get good understanding of user sentiment on antidepressants and finding influential users, and reacts on drug side effects. Our work is organized as follows. The related and literature work depicted under Section II and

the proposed work and its contribution described under Section III and results and discussions under Section IV and conclusions is depicted under Section V.

2. Literature Study

The proposed work is inspired by the work that is done [1], where the authors contributed the below methods to find antidepressants. The overall scenario is shown in Figure 1:

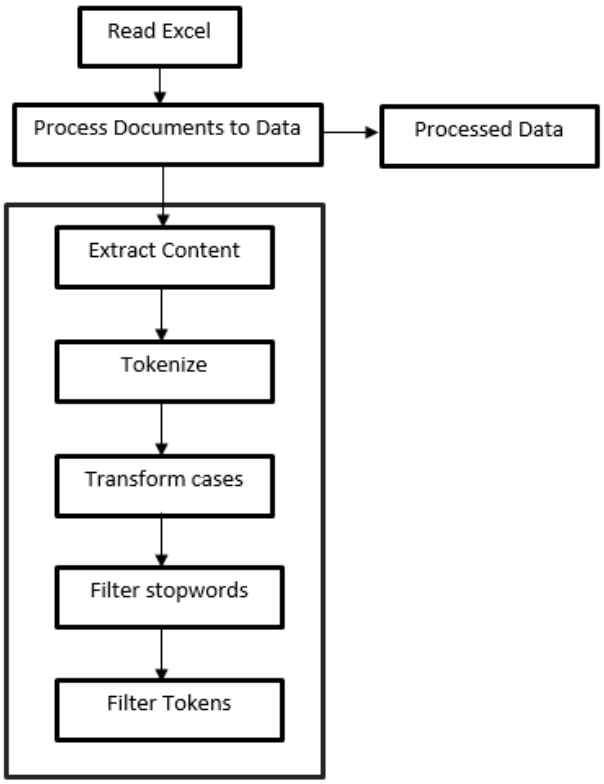

Fig 1: Processing of Rapid miner to get TF-IDF Scores.

The initial step was a search for depression dedicated forums. Our last list, which produces the below chart of descending order. Hence we choose depressionforums.org. After that collecting, analyzing and processing tree of data was generated in Rapid miner(www.rapidminer.com) to detect the most happening words(positive, negative and side effects) to get their Term-Frequency-Inverse Document Frequency) scores within each post. The Figure 1 displays the collection and processing of a tree. The dataset was uploaded ("Read Excel"), processed ("Process Document to data") with the use of subcomponents("Extract Content", "Tokenize", "Transform Cases", "Filter Stop words", "Filter Tokens" respectively) that filters extra noise (misspelled words, common stop words etc.) to have variable measures uniformity. The output ("Processed Data") has the final word list; with each word has a particular TF-IDF score. The TF-IDF scores in every post that built was dependent on a representative word set in entire forum and reflects the semantic posts content. Hence, we showed a TF-IDF vector as every post's semantic profile.

Significantly, much similarity measure can be derived to show how close the two post's semantic profiles are, like Euclidian distance or correlation. In addition to this cluster analysis will performs to find the groups of same semantic profiles. They used *k*-means clustering [11] to group all the posts semantic profiles of our forum as a mandatory preprocessing step for network based modeling. Later the further network based modeling step of posting forum is applied. The activity of forum posting having of threads with thousands of postings and responses were designed as a high user centric network. This modeling approach aiming at displaying user interactions by taking posts semantic content into consideration. Our network nodes correspond to forum users and connects directed edges correspond to direct and context interactions. The user-to-user replies with the use of forum's "Reply" option. Those interactions are said to be direct interactions and are modeled with the edges of bidirectional connects the corresponding nodes. This made us to mutual data exchange between a poster and a direct replier. The users posting within a particular thread (threads may be topic related and thread semantic content is same).are reflected by context interactions. The network nodes sample is shown in figure 2.

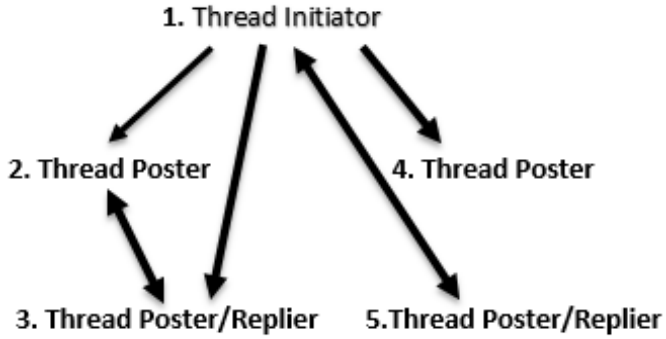

Fig 2: Sample Network Model with Edges.

Significantly, the forum posts are converted into a huge directional weighted network with multiple densely connected units (or modules). The network partitioning algorithm helps in detecting the nodes' over representation within every module. For this objective we used HITS, which is a initially developed method for web pages link analysis [12, 13]. Data disseminates from authoritative nodes. There is a link from hubs to authoritative nodes and so they broke flow of information inside the network. This approach helps in finding influential users and benefit in considering the structural properties of both networks' and the direction of data flow.

Lise Getoor *et al*. took more datasets which are depicted as linked collections of objects which are interrelated. These represent homogeneous type networks with single object type and link type and heterogeneous networks with multiple object and link types (other semantic data if possible). Single mode social networks like people connect by friendship links or WWW, a linked web pages collection are the examples of homogeneous networks whereas medical domains depicting patients, diseases, treatments and contacts, authors and so on are the instances of heterogeneous networks.

Link mining is one of data mining techniques that externally take these links when developing predictive or descriptive models of linked data. In general link mining activities include ranking objects, finding groups, collective classification, link prediction and discovery of sub graph. Where network analysis was studied in specified areas like analysis of social

network, hypertext mining and web analysis only in recent there have been a ideas cross-fertilization among those varied communities. This is an emerging area. In this paper we reviewed some common emerging themes. The recent considerable interests in algorithms were proposed by M.E Newman *et al*. for detecting the communities in networks. The connections within the groups of vertices are denser while the connections between them are sparser. The progresses reviewed here which is tending towards this end. We start by depicting some conventional methods in finding community like spectral bisection, the Kernighan Lin algorithm and hierarchical clustering on similar measures basis. Not any method is same for real-world network data kinds with which present research concerns like Internet and web data and biological and social networks.

Kleinberg HITS and the Google PageRank algorithms was proposed by Andrew Y. Ng, Alice X. Zheng *et al*. Which are eigenvector methods for detecting "authoritative" or "influential" articles, given information on hyperlink or citation? Those algorithms must give reliable answers is exactly a desideratum and in [10], we analyzed when they are expecting stable rankings under small perturbations to patterns linkage. In this paper we extend the analysis and shows how it gives insight into ways of stable links designing analysis methods which in turns motivates two more new algorithms using citation and web hyperlink data.

The social network's snapshot was given by David *et al*., could we infer which recent interactions are likely going to occur in future? We formalize the above question as the link expected issue, and generate approaches to link expected that measure for network nodes proximity. Experiments on large co-authorship network assist that data regarding future interactions can be gathered from network topology alone and that fairly measures the proximity of node can outperform more direct measures.

Felix Naumann *et al*., took more biological data sources that has data on scientific entities classes like genes and sequences. The scientific object's logical relationships are implemented like URL's and foreign ID's. To traverse links and paths (links concatenation) through these sources Query processing is performed. We design data objects in these sources and an object graph is the links between objects. We detect a collection of interesting properties for links and paths like out degree, link image, data object's cardinality and links, the number of distinct objects reached by some links and so on. Analogous to database cost models; To develop a framework from object graph we used statistics for estimating query result size on object graph. Analogous to training and testing, to estimate the result size we used sampled data from queries. Our models are validated with the use of sample data from NIH/NCBI data sources. Our research provides a foundation data sources querying and exploring.

Jan Noessner *et al*. was argued that linked open data is the major advantage of semantic technologies for web since it gives more structured information with effective way of access than web pages. In this paper, we proposed a new approach for object recognition that relies on prevailing semantic similarity measure for linked data. We choose a measure to the problem in object recognition, and presented precise and relevant algorithms that implement the methods and give a systematic experimental evaluation on benchmark dataset basis. As our result, we shown that the use of lightweight ontologies and schema information mainly enhances object recognition in the terms of linked open data.

3. Proposed Work

This enabled us to represent exact user interactions by depending on data's semantic content. In this paper, it enabled us for precise representations of user interactions with the dependency on data's semantic content. Analyze the user interactions and extract extra knowledge of user's posts. Our work key advantage is increasing health solutions for patients with depression and the person who is affected with depression can also share their experiences.

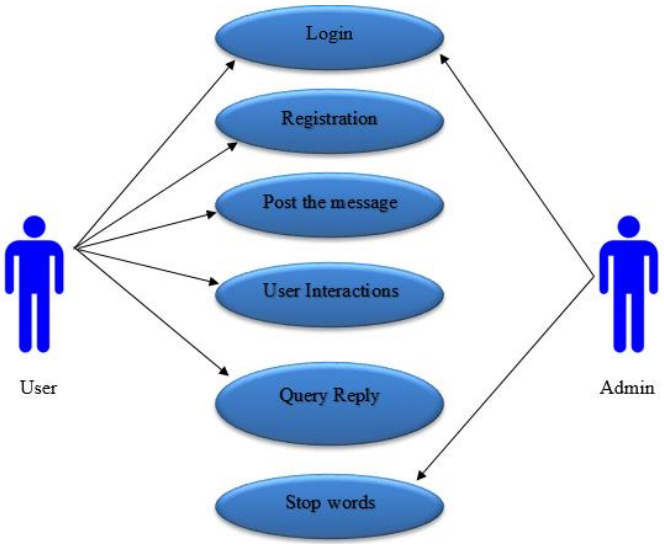

Fig 3: Use case diagram of our work

We analyzed content network site pages and search engines commit to allot a category for page. A Software chooses the category from finite list. Likewise the keywords and ad text of content campaign are analyzed and every ad group will be given a category of the same list. We used Semantic search seeks to enhance accuracy of search by knowing the searcher's intension and the terms context as they are shown in data space that is searchable, either on the Web or inside a closed system, for generating some more related outcomes. The work contributes in 5 steps:

1. User Interface Design

The user's important role is moving login window to social network. It is done for the purpose of security. In login page the login user id and password are entered. It checks for username and password are matched or not (valid user id and password).If user name and password are not valid we are unable to enter login window and it will displays error message to data owner window. Thus we prevent the login window to social network from unauthorized access. It made our project more secure. So server has user id and password and server also performs user authentication. It enhances the security and prevents illegal access by the user to enter the network. For our work JSP is used for designing. We validated the user's login and authentication of server.

2. User Upload Posts

Social media, from personal texts to live fora's providing boundless chances to users to share their experiences. Additionally it is providing more chances for companies for getting feedback on their products and services. Many companies are now concentrating on social network monitoring as their first priority in their IT departments, and also creating a chance for getting rapid feedback on their products and services to reduce and improve delivery, increase turnover and profits and minimize costs.

3. Admin Analyzing Posts

On the basis of user opinions on the posts the structures are determined by the initial exploratory analysis. The outputs are the user's clusters compilation and their opinion on the posts. To determine the users who are influenced among the members subsequent analysis was used.

4. Admin Block Posts

A multi consensus relates to admin relies on every patient solutions. The platform of social media results in individuals with varied outputs depending on different individual factors and circumstances. Apart from those factors we are able to move through the data and can collect favorable and unfavorable sentiment, which was then decided by research that emerged on user's effectiveness.

5. User View Posted Information

Whenever the user login the interface and they share the opinion in Forum either positive or negative. Most of the users may login the forum and can share their experience. The user who is influenced can reveal their opinion about specific topics. It is beneficial for being aware about general information. The overall architecture of our work is shown in Figure 4.

4. Results and Discussion

We have concentrated to reduce the content which is not matched with the topic. Like using abuse words or words those are irrelevant are monitored and can't be used because we considered this as stop word. For instance, let fora be the application regarding depression disease. Heartache, stomach pains are the words treated as irrelevant and abusive so that specific medicine or a person will be strictly controlled by this approach. Firstly the user will login to fora with authenticated details. If won't he can provide details for registration. The home screen has the posts lists which are shared by different people on specific drug or on disease. They are also able to assist the disease symptoms and better treatment for that disease, good available hospitality with the address locations or any post related images. The users can reply to the post by sharing the other's view on the before posts. To avoid misleading communication among users stop words technique will be used. Any viewer can view the post and can like or dislike the posts. The number of likes or votes is helpful in the users' group identification that those are interested in or faced the same experience. We show some of our work related screenshots in the following figures. Figure 5 shows the admin screen, where user is able for adding stop words into prevailing list.

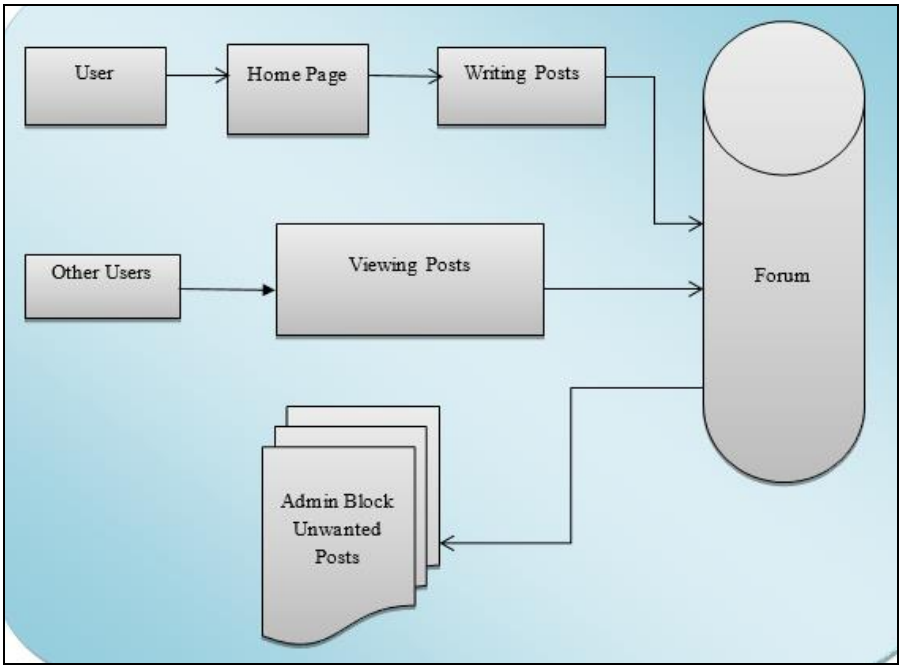

Fig 4: System Architecture

Once the user added the word into database, the word is unavailable for any user. Here I was added the word heartache into the list, because we took this fora related to depression**.**

heartache
Python
C++
Java

Fig 5: Sample list of stop words.

The below figure is our website's home page. The description of the post is stored in the centered followed by the posted person and with the number of views; votes and any other reply proceeded to a particular post. For instance, a reply for first post is shown in figure 7 with the corresponding replies.

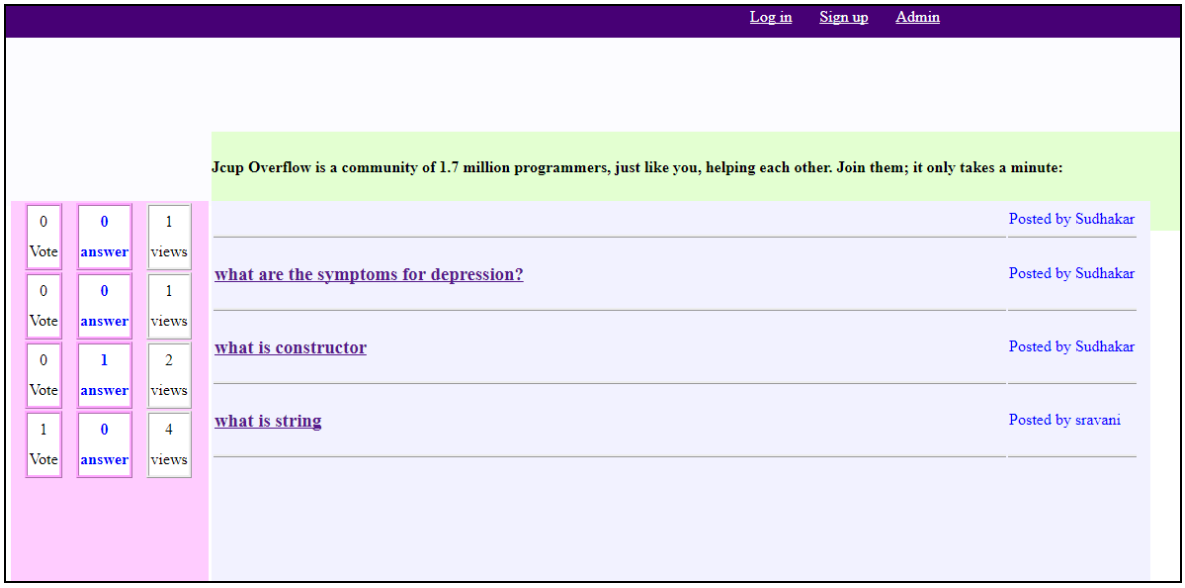

Fig 6: Home Page of our forum

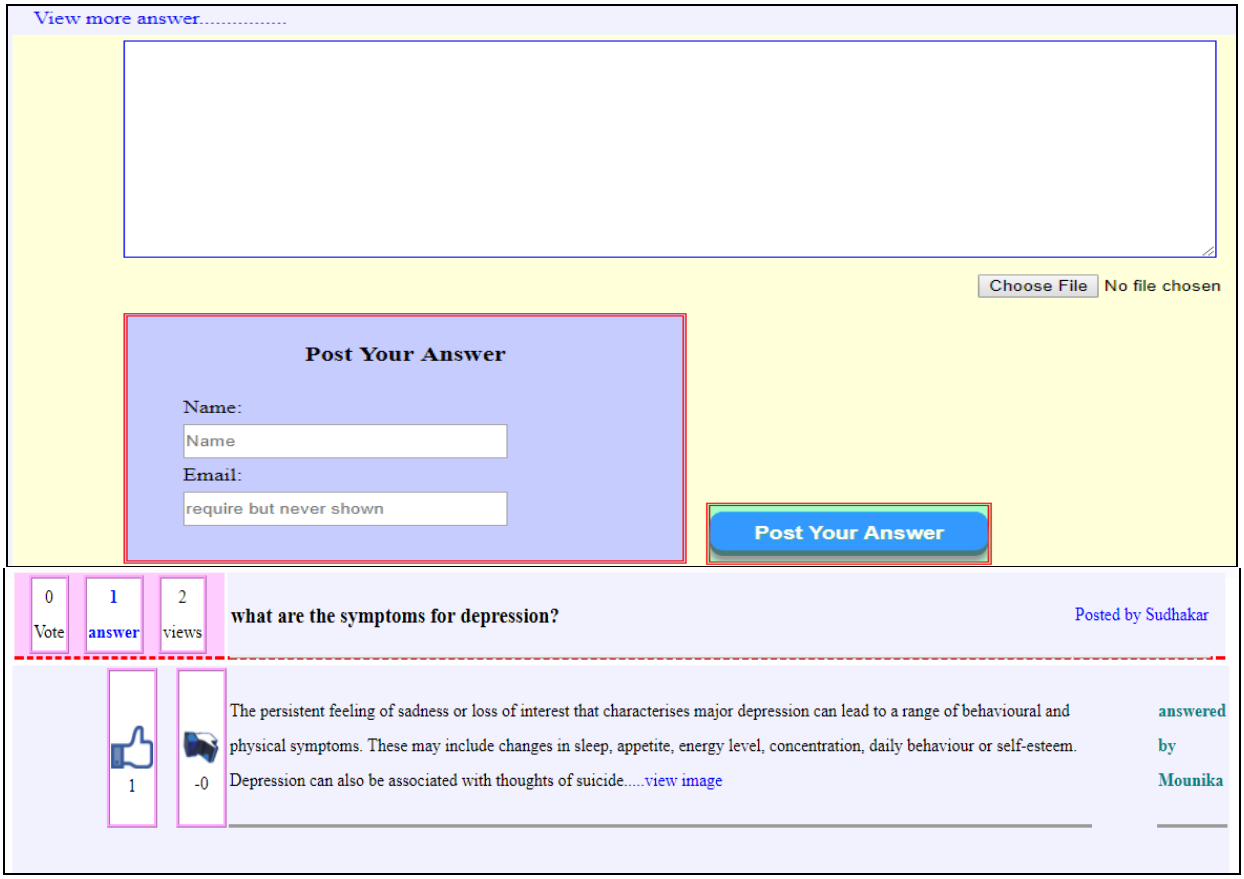

Fig 7: User Interaction for the post

5. Conclusion

Depression is said to be the main reason for disability, a wellbeing and burden of various diseases in global range and 300 million people are affecting in world wide. The depression that can't be solved will leads to various issues from stroke to coronary diseases, both of main illness and cause death in 2013. A virtual system as social network that is composed with nodes and edges has more content. Those

contents can be designed and gathered using various tools that are trendy and formulate expectations and builds the relationships among users. Graphical portrayal provides the information clearly on these user interactions. In this paper we contributed 3 aspects. One is the user activity monitoring and followed by network clustering ad module analysis. An individual who likes a particular post will be taken as a group while the other belongs to other group. To avoid misleading communication among users stop words technique is used and also for efficient interaction of user. The statistical analysis of such user interactions is beneficial in health networks to be aware of particular diseases. It enables us all the gatherings took part and for enhancements in healthcare to suffering people from disease in future. Eventually concluding that using these types of data mining systems can widely enhance the healthcare system's quality at cheaper cost with in time.

6. References

1. World Health Organization. (2015, Jan. 5). Depression Fact Sheet No.3 69. [Online]. Available:http://www.who. int/mediacentre/factsheets/fs369/en/
2. Altug Akay *et al*. Assessing Antidepressants Using Intelligent Data Monitoring and Mining of Online For a in IEEE Journal of Biomedical and Health Informatics, 2016, 20(4).
3. Ferrari AJ, *et al*., Burden of depressive disorders by country, sex, age, and year: Findings from the global burden of disease study, PLOS Med., 2010-2013.
4. Web MED. (2015, Jan. 6). Depression Health Center: Untreated Depression. [Online]. Available: http://www.webmd.com/depression/guide/untreated-depression-effects
5. World Health Organizations, 2015. The Top 10 Causes Death. [Online]. Available: who.int/media Centre/facts heets/fs310/en/
6. Wu JL, Yu LC, Chang PC. Detecting causality from online psychiatric tests using inter-sentential language patterns, BMC Med. In format. Decision Making, 2012, 12.
7. Moradi F, Eklund AM, Kokkinakis D, Olovsson T, Tsifas P. A graph-based analysis of medical queries of a swedish health care portal, in Proc. 5th Int. Workshop Health Text Mining Inf. Analysis, Gothenburg, Sweden. 2014; 26(30):2-10.
8. Bedmar IS, Revert R, Martinez P. Detecting drugs and adverse events from Spanish health social media streams, in Proc. 5th Int. Workshop Health Text Mining Inf. Analysis, Gothenburg, Sweden. 2014; 26(30):106-115.
9. Paul M, Dredze M. Discovering health topics in social media using topic models, PLOS One. 2014; 9:e103-e408.
10. Zhao K, *et al*. Finding influential users of online health communities: A new metric based on sentiment influence, J Amer. Med. Inform. Assoc. 2014; 21:212-218.
11. Kanungo T, *et al*., An efficient k-means clustering algorithm: Analysis and implementation, Pattern Anal. Mach. Intell. Trans. 2002; 24(7):881-892.
12. Kleinberg JM. Authoritative sources in a hyperlinked environment, J ACM. 1999; 46(5):604-632
13. Pujol JM, Sanguesa R, Delgado J. Extracting reputation in multi agent systems by means of social network to Pology, inProc.1stInt.Joint Conf. Auton. Agents Multi agent Syst, 2002, 467-474.