IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 6, Ver. III (Nov – Dec. 2015), PP 01-09
www.iosrjournals.org

Sentiment analysis for improving healthcare system for women

Jeyalakshmi Jeyabalan[1], Sindhuja Manivasagam[2], Poornimathi Krishnan[3], Sreesubha Soundarrajan[4], Anitha Jaikumar[5]
[1,2,3,4,5]*Department of Information Technology Rajalakshmi Engineering College Thandalam, Chennai, India*

Abstract: *The system proposes a feedback mechanism wherein, sentiment analysis is performed from surveys and tweets based on prevailing health issues among adult women in India and the social opinion on prevalent health issues is analyzed, and measures are taken to create awareness using email, SMS, blog, forum posts or web site posts. The system focuses on study of opinions and subjects discussed in the forum. Sentiment analysis is performed on this genre and if positive emotions are asserted, then awareness programs can be initiated for Thyroid issues and Stress Control. Periodically current issues are initiated and Sentiment analysis is performed, consequently awareness initiatives are created. Thereby, helping the performance improvement of healthcare initiatives. This awareness initiative helps concentrating on current health issues that are widespread and if the reach of such awareness programs is better, then the awareness created may have high impact amid the middle aged women.*

Index Terms: *Sentiment Analysis, Pervasive Healthcare*

I. Introduction

India currently has one of the highest rates of malnourished women among developing countries. A study in 2000 found that nearly 70 percent of non-pregnant women and 75 percent of pregnant women were anaemic in terms of iron-deficiency. One of the main drivers of malnutrition is gender specific selection of the distribution of food resources. Therefore, national health programs such as the National Rural Health Mission (NRHM) and the Family Welfare Program have been created to address the maternal health care needs of women across India. Maternal malnutrition has been associated with an increased risk of maternal mortality and also child birth defects. Addressing the problem of malnutrition would lead to beneficial outcomes for women and children.

Furthermore it was found that families failed to seek medical treatment for their daughters because of the stigma associated with negative medical histories. A study conducted by Pednekar et al. in 2011 found that out of 100 boys and girls with congenital heart disease, 70 boys would have an operation while only 22 girls will receive similar treatment.

In the present situation, it is mandatory that serious measures be taken for building an awareness campaign amid adult women in India. The advent of a wide spread mobile network even in deep rural areas of India and the availability of Internet over phones at cheaper costs give a strong helping hand to build such a healthcare initiative to educate women and enlighten families. For women who can't express their health issues to others, such portals can help a lot. The proposed system can also keep abreast with current health issues women are dealing with and addressing them by building awareness amid them related to the disease.

The paper presents the related work on chapter 2, system description and architecture on chapter 3 and discussion of results on chapter 4.

II. related work

In India, it is commonly observed that Women deny themselves and don't easily take care of themselves. To build an efficient and effective health system, it is mandatory that health related awareness be created amid women who may also educate their children and family members. Women's health in India can be also an important contributor to economic growth because these days most women also contribute to family income by taking a job. Currently, women in India face a multitude of health problems, which ultimately affect the aggregate economy's output. Due to the social structure, Indian women are more likely to have difficulty travelling in public spaces than men, resulting in greater difficulty to access services.

India currently has one of the highest rates of malnourished women among developing countries. Although India has witnessed dramatic growth over the last two decades, maternal mortality remains stubbornly high in comparison to many developing nations As a nation, India contributed nearly 20 percent of all maternal deaths worldwide between 1992 and 2006. The primary reasons for the high levels of maternal mortality are directly related to socioeconomic conditions and cultural constraints limiting access to care.

A 2012 study by Tarozzi have found the nutritional intake of early adolescents to be approximately equal .However, the rate of malnutrition increases for women as they enter adulthood. Furthermore, Jose et al. found that malnutrition increased for ever-married women compared to non-married women.

India is facing a growing cancer epidemic, with a large increase in the number of women with breast cancer. By the year 2020 nearly 70 percent of the world's cancer cases will come from developing countries, with a fifth of those cases coming from India. As of 2012, India has a shortage of trained oncologists and cancer centres, further straining the health care system.

Cardiovascular disease is a major contributor to female mortality in India. Indians account for 60% of the world's heart disease burden, despite accounting for less than 20% of the world's population. Indian women have a particular high mortality from cardiac disease and NGOs such as the Indian Heart Association have been raising awareness about this issue. Women have higher mortality rates relating to cardiovascular disease than men in India because of differential access to health care between the sexes. One reason for the differing rates of access stems from social and cultural norms that prevent women from accessing appropriate care. For example, it was found that among patients with congenital heart disease, women were less likely to be operated on than men because families felt that the scarring from surgery would make the women less marriageable.

Mental health consists of a broad scope of measurements of mental well being including depression, stress and measurements of self-worth. Numerous factors affect the prevalence of mental health disorders among women in India, including older age, low educational attainment, fewer children in the home, lack of paid employment and excessive spousal alcohol use. One of the most common disorders that disproportionately affect women in low-income countries is depression. Indian women suffer from depression at higher rates than Indian men. Indian women who are faced with greater degrees of poverty and gender disadvantage show a higher rate of depression. The difficulties associated with interpersonal relationships—most often marital relationships—and economic disparities have been cited as the main social drivers of depression.

The above mentioned facts mandate that serious measures have to be taken for building an awareness campaign amid adult women in India. The advent of a wide spread mobile network even in deep rural areas of India and the availability of Internet over phones at cheaper costs give a strong helping hand to build such a healthcare initiative to educate women and enlighten families. For women who can't express their health issues to others, such portals can help a lot. The system can also keep abreast with current health issues women are dealing with and addressing them by building awareness amid them related to the disease.

The system proposes a feedback mechanism wherein, sentiment analysis is performed from surveys and tweets based on prevailing health issues among adult women in India and the social opinion on prevalent health issues is analyzed, and measures are taken to create awareness using email, SMS, blog, forum posts or web site posts. The system focuses on study of opinions and subjects discussed in the forum. The system's working is briefed with an example. The genre for sentiment analysis is considered as below for example. A lot of adult women around 30 years in India have been observed with Thyroid issues in recent years, which is largely contributed by stress factor. Sentiment analysis is performed on this genre and if positive emotions are asserted, then awareness programs can be initiated for Thyroid issues and Stress Control. Periodically current issues are initiated and Sentiment analysis is performed, consequently awareness initiatives are created. Thereby, helping the performance improvement of healthcare initiatives. This awareness initiative helps in concentrating on current health issues that are widespread and if the reach of such awareness programs is better, then the awareness created may have high impact amid the middle aged women.

Sentiment analysis in healthcare uses natural language software to categorize and assess written and spoken comments by patients about their healthcare experience and ideas. Sentiment analysis provides healthcare organizations with much deeper insight into patient perceptions and an understanding of where changes can have the most dramatic impact on improving healthcare performance improvement.

Sentiment analysis (also known as opinion mining) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. Generally speaking, sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document. The attitude may be his or her judgment or evaluation (see appraisal theory), affective state (that is to say, the emotional state of the author when writing), or the intended emotional communication (that is to say, the emotional effect the author wishes to have on the reader).

A basic task in sentiment analysis is classifying the polarity of a given text at the document, sentence, or feature/aspect level — whether the expressed opinion in a document, a sentence or an entity feature/aspect is positive, negative, or neutral. Advanced, "beyond polarity" sentiment classification looks, for instance, at emotional states such as "angry," "sad," and "happy."

Another research direction is subjectivity/objectivity identification. This task is commonly defined as classifying a given text (usually a sentence) into one of two classes: objective or subjective. This problem can sometimes be more difficult than polarity classification: the subjectivity of words and phrases may depend on their context and an objective document may contain subjective sentences (e.g., a news article quoting people's opinions). Moreover, results are largely dependent on the definition of subjectivity used when annotating texts.

A more fine-grained analysis model is called the feature/aspect-based sentiment analysis. It refers to determining the opinions or sentiments expressed on different features or aspects of entities, e.g., of a cell phone, a digital camera, or a bank. A feature or aspect is an attribute or component of an entity, e.g., the screen of a cell phone, or the picture quality of a camera. This problem involves several sub-problems, e.g., identifying relevant entities, extracting their features/aspects, and determining whether an opinion expressed on each feature/aspect is positive, negative or neutral.

The proposed system goes for sentiment analysis and to the fine grained level also measures the subjectivity of the text under study.

III. System Description And Architecture

Sentiment Analysis has been performed with several methodologies and resources. The word space models were created using following techniques Latent Semantic Analysis (document based co-occurrence), Hyperspace Analogue to Language (word based co-occurrence) , Latent Dirichlet Allocation or Random Indexing. It is found that a word space model that is inherently incremental and does not require a separate dimension reduction phase using random indexing. Similarly SVM have been proven as one of the most powerful learning algorithms for text categorization. The literature survey for the considerations performed are presented below.

Ms. Gaurangi Patil[1] et al., proposed a Sentiment analysis where the classifiers for sentiment analysis of user opinion towards political candidates through comments and tweets using Support Vector Machine (SVM) is presented. Classifier is developed that performs sentiment analysis, by labeling the users comment to positive or negative from which we can classify text into classes of interest.

Balamurali[2] et al., proposes a semantic features using word senses for a supervised document-level sentiment classifier.To highlight the benefit of sense-based features, word-based representation of documents and a sense-based representation are compared where WordNet senses of the words are used as features. Finally, it is shown that even if a WSD engine disambiguates between a limited set of words in a document, a sentiment classifier still performs better than what it does in absence of sense annotation. Since word senses used as features show promise, the possibility of using similarity metrics defined on WordNet are also examined to address the problem of not finding a sense in the training corpus. Experiments are performed using three popular similarity metrics to mitigate the effect of unknown synsets in a test corpus by replacing them with similar synsets from the training corpus.

Aditya Joshi [3] et al., propose Random Indexing as a simple implementation of Random Projections with a wide range of applications. It can solve a variety of problems with good accuracy without introducing much complexity. It is demonstrated for identifying the language of text samples, based on a novel method of encoding letter n-grams into high-dimensional Language Vectors. Further, it is shown that the method is easily implemented and requires little computational power and space. As proof of the method's statistical validity, we show its success in a language-recognition task.On a difficult data set of 21,000 short sentences from 21 different languages, and system achieved 97.8% accuracy, comparable to state-of-the-art methods.

Tanveer Ali[4] et al., proposed Text mining studies that have started to investigate relations between positive and negative opinions and patients' physical health. It shows, Several studies linked the personal lexicon with health and the health-related behavior of the individual. However, few text mining studies were performed to analyze opinions expressed in a large volume of user-written Web content. The current study focused on performing sentiment analysis on several medical forums dedicated to Hearing Loss (HL). The system categorized messages posted on the forums as positive, negative and neutral. T he study had two stages: first, manual annotation of the posts is applied with two annotators and have 82.01% overall agreement with kappa 0.65 and then Machine Learning techniques to classify the posts are applied.

Yihan Deng[5] et al., ponder over the Physicians and nurses expressing their judgments and observations towards a patient's health status in clinical narratives. Their judgments are explicitly or implicitly included in patient records. To get impressions on the current health situation of a patient or on changes in the status, analysis and retrieval of this subjective content is crucial. In this paper, the authors approach this question as sentiment analysis problem and analyze the feasibility of assessing these judgments in clinical text by means of general sentiment analysis methods. Specially, the word usage in clinical narratives and in a general text corpus is compared. The linguistic characteristics of judgments in clinical narratives are collected. Besides, the requirements for sentiment analysis and retrieval from clinical narratives are derived.

Bo Pang[6] et al., presents a survey covering techniques and approaches that promise to directly enable opinion-oriented information-seeking systems. The focus is on methods that seek to address the new challenges raised by sentiment-aware applications, as compared to those that are already present in more traditional fact-based analysis. It includes material on summarization of evaluative text and on broader issues regarding privacy, manipulation, and economic impact that the development of opinion-oriented information-access services gives rise to. To facilitate future work, a discussion of available resources, benchmark datasets, and evaluation campaigns is also provided.

Despo Georgiou[7] et al., propose a project aimed to examine a number of tools regarding their suitability for healthcare data. A comparison between commercial and non-commercial tools was made using responses from an online survey which evaluated design changes made to a clinical information service. The commercial tools were Semantria and TheySay and the non-commercial tools were WEKA and Google Prediction API. Different approaches were followed for each tool to determine the polarity of each response (i.e. positive, negative or neutral). Overall, the non-commercial tools outperformed their commercial counterparts. However, due to the different features offered by the tools, specific recommendations are made for each. In addition, single-sentence responses were tested in isolation to determine the extent to which they more clearly express a single polarity. Further work can be done to establish the relationship between single-sentence responses and the sentiment they express.

Wingyan Chung[8] et. al., proposed a framework for designing business intelligence systems that extract the relationship between the customer ratings and their reviews. This framework is developed based on algorithms such as inductive rule learning, rough set theory (RST), and information retrieval. It is experimented with the association rule mining (ARM) method, RST exhaustive algorithm, and RST LEM2 algorithm. Among three algorithms, ARM algorithm achieved the highest level of support and highest confidence for the product with the largest number of reviews, while RST algorithms produced rules that have the highest confidence values.

Khairullah Khan [9] et. al., presented a systematic literature survey regarding the computational techniques, models and algorithms for mining opinion components from unstructured reviews. This survey mainly focused on subjectivity analysis and lexical resource generation.

Arturo Montejo-Ráez[10] et. al.,proposed an unsupervised approach to the problem of polarity classification in Twitter posts.This approach uses Personalized Page Rank vectors (PPVs) for expanding senses , which is to represent each tweet as a vector of weighted synsets that are semantically close to the terms included in the post.PPV is a ranked sequence of WordNet synsets weighted according to a random walk algorithm. It includes two main resources such as a graph to connect terms and polarity weights for individual terms.

Shoushan Li[11] et. al., proposed an machine learning approach to incorporate polarity shifting information into a document-level sentiment classification system.Training data is produced by feature selection method to generate binary classifier. Two classifier combination methods are applied to perform polarity classification.

In order to have an efficient classification of text data Support vector machines and random indexing are used. With the increase in dimensionality, random indexing and Support Vector Machines prove to be better in performance.

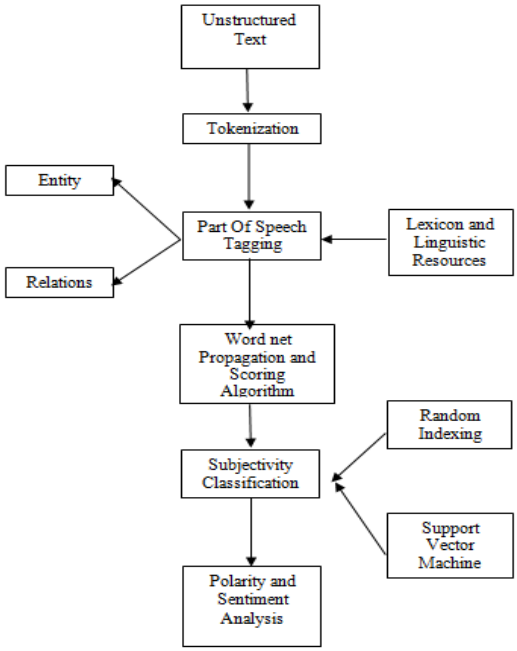

Fig 1: System Architecture

The Text that is under test is tokenized as shown in Fig 1, speech tagging and subject orientation is performed, then using wordnet and synset, a sentiwordnet is formulated, then it is interpreted and sentiment orientation is performed. In order to make this happen the following resources are used,

1. Bing Liu's Opinion Lexicon
2. MPQA Subjectivity Lexicon
3. SentiWordNet
4. Harvard General Inquirer
5. LIWC - Linguistic Inquiry and Word Counts

The above resources help to formulate the Sentiwordnet for the test data using Wordnet Propagation and Scoring algorithms. Co - Occurence Matrices are formed from the words in the document. Random indexing is used to create a word space model. Polarity is identified from these methods. Support Vector Machines are used to classify the text of interest.

The main objectives of the system are discussed below. To analyze the opinions on prevalent health issues amid adult women in India and initiate awareness programs based on the issue discussed is the main objective of the system. The opinions and subjects discussed are considered for analysis to find possible best practices or indications towards what types of healthcare initiatives need to be made available among adult women in India. The analyzed results are used to create healthcare content and can be hosted as tweet or web posts. This content viewed by the end users i.e middle aged women especially working women, can create huge change in the healthcare viewpoint of the women. This can serve as a good source for understanding and addressing the current health issues which many state agencies do not even address presently. The analyzed information can be disseminated to rural segments of India by means of Healthcare officials and volunteers. The proposed system satisfies following objectives.

1. Individual subject's disease pattern is identified.
2. The disease pattern of a group of people taking same treatment is identified.
3. The current healthcare issues are identified and categorized from subjectivity of the free text or comments.
4. The results are used to create content for creating awareness among the adult women using a healthcare forum as a website after expert consultation.
5. The root causes and minor criteria like stress for example indirectly causing major diseases are identified as part of the analysis.
6. Comments by patients are categorized as either positive or negative descriptions of their health care.
7. The system automatically predicts whether a patient would recommend a therapy or a treatment.

IV. Results and discussion

The results of the proposed system and possible impact is discussed as below. Sentiment analysis is performed using Sentiment viz tool and visualization of the tweets over any genre is analyzed. The tag cloud is

formed and sentiment of the tweets are generated. The tag cloud indicates the most discussed topic with larger font size and others in decreasing order of their sizes. The most prominent issues are solutions are taken for content creation and dissemination over social media. The results are discussed as below.

The following figures Fig 2 and Fig 3 are the results of Sentiment Analysis over Diabetes in Women. Fig 2 indicates the sentiment of the tweets. Fig 3 indicates the tag cloud formed by inferring the tweets. The tag cloud indicates that more is told over Gestational Diabetes, Weight Loss, Type of Diabetes, Diet over Diabetes etc.,

The figures Fig 4 and Fig 5 are the results of Sentiment Analysis over Thyroid in Women. Fig 4 indicates the sentiment of the tweets. Fig 5 indicates the tag cloud inferred from the tweets. The tag cloud indicates that Pregnancy, Menopause, 30s Age group, Hormonal issues, obesity and iodine disorders, cancer related facts etc.,

The chart displayed on Fig 6 indicates the trend of tweets over Diabetes in Women. For 100 tweets on the genre, it is seen that the tag cloud indicates Gestational Diabetes, Weight Loss, Menopause, Age group of women are the most prominent issues over others. The online tweet2scv site fetches and converts the tweets to csv file. Tag cloud is formed from the content and the tags and their frequencies are summarized in Fig 6.

Thus it is possible to identify the current trend of the genres discussed for example gestational diabetes in women and thereby more awareness programs are created to prevent the diseases.

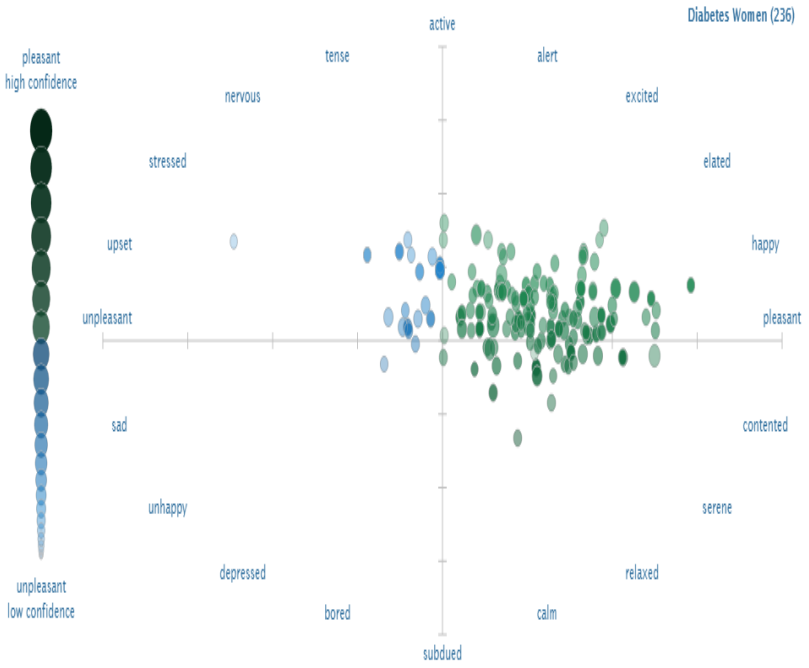

Fig 2: Sentiment Analysis for "Diabetes in Women"

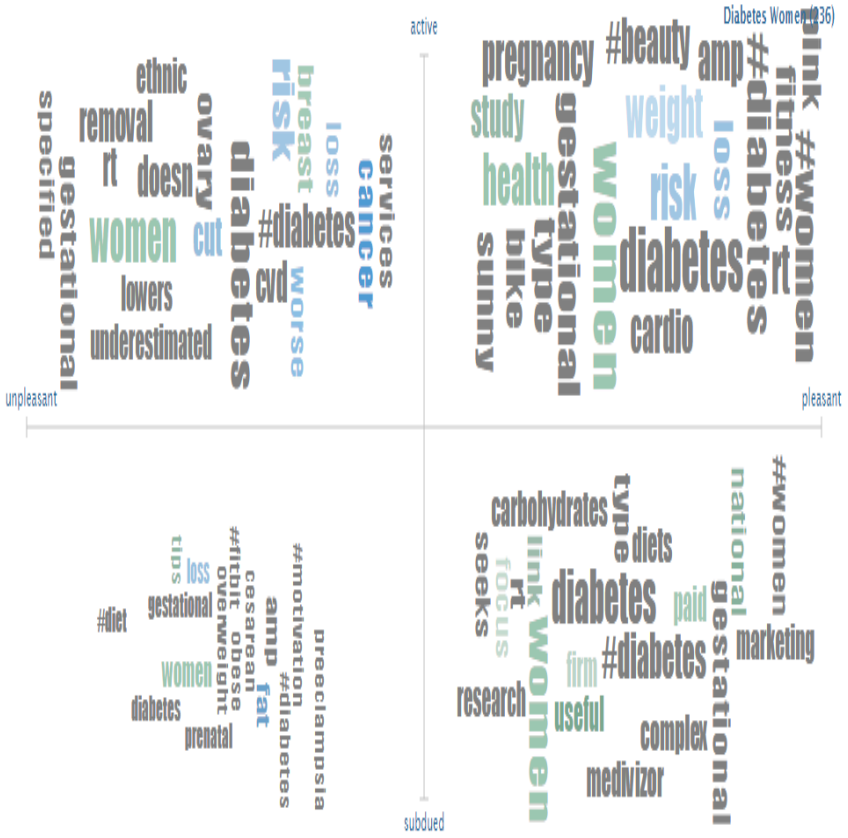

Fig 3: Tag Cloud for "Diabetes in Women"

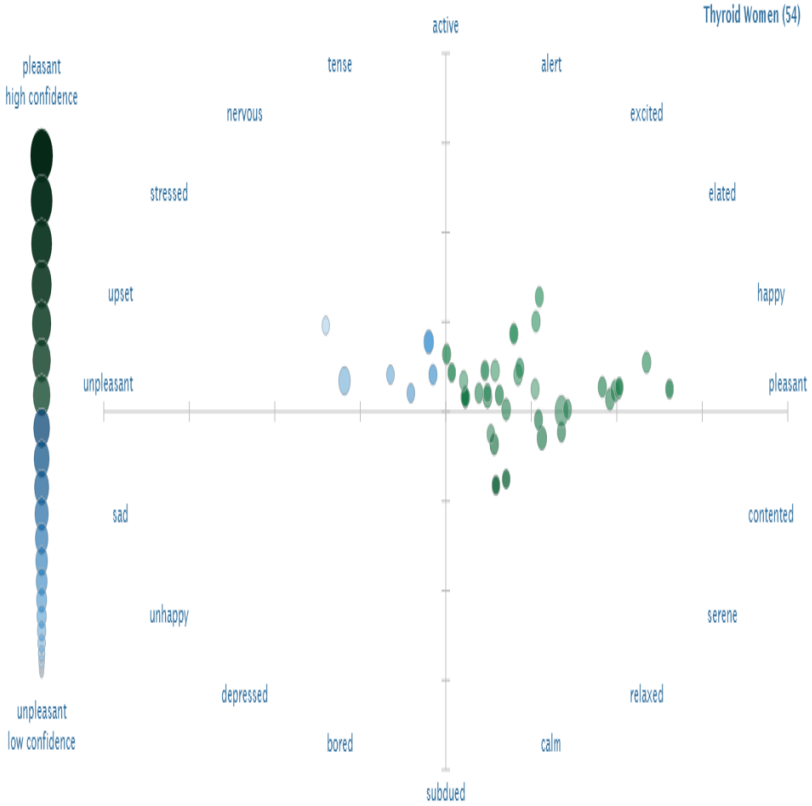

Fig 4: Sentiment Analysis for "Thyroid in Women"

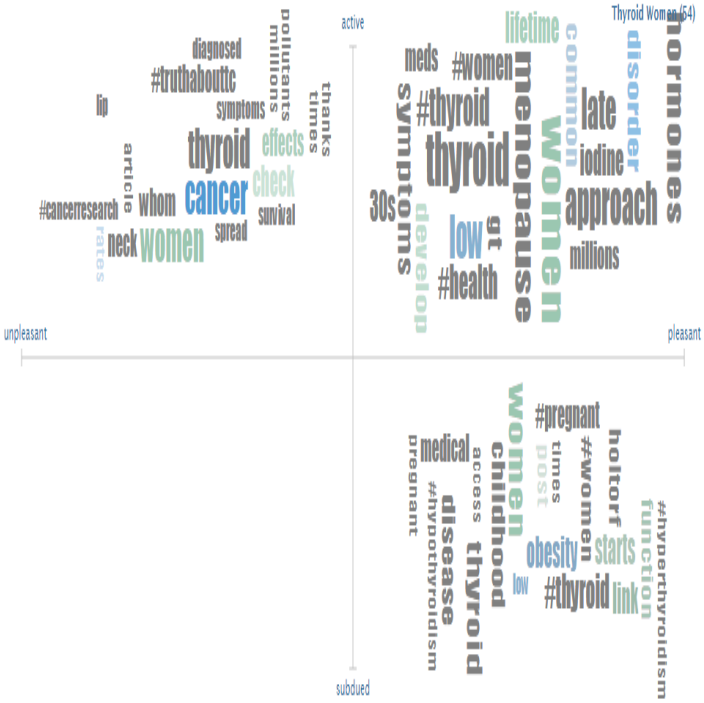

Fig 5: Tag Cloud for "Thyroid in Women"

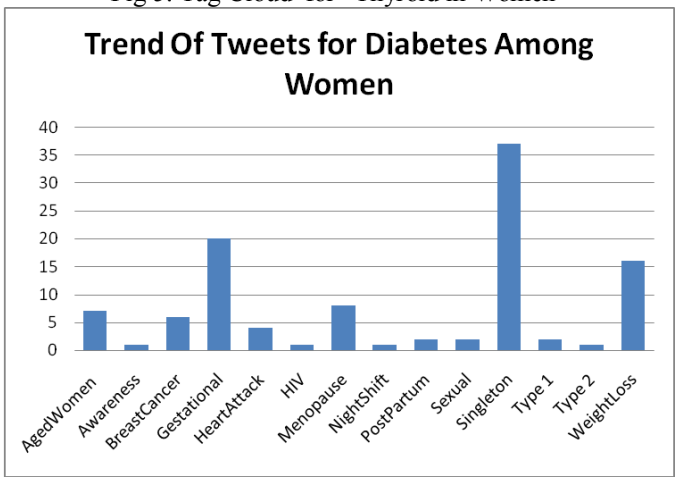

Fig 6: Trend Analysis of Tweets "Diabetes in Women"

V. Conclusion

Thus a feedback mechanism with sentiment analysis is performed from surveys and tweets based on prevailing health issues among adult women in India and the social opinion on prevalent health issues is analyzed, and measures are taken to create awareness using email, SMS, blog, forum posts or web site posts. The results show the implementation of twitter sentiment analysis and analysis of the tweets, sentiment of the tweets and the tag cloud inferred from there for implementation of awareness programs. This system can prove to be a vital and indispensable utility for reducing the mortality and morbidity rates due to gender specific diseases.

References

[1] Ms. Gaurangi Patil, Ms. Varsha Galande, Mr. Vedant Kekan, Ms. Kalpana Dange, "Sentiment Analysis Using Support Vector Machine", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 2, Issue 1, January 2014, pp - 2607-2612

[2] Balamurali A R, Aditya Joshi Pushpak Bhattacharyya, "HarnessingWordNet Senses for Supervised Sentiment Classification", Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, UK, July 27–31, 2011, pp- 1081–1091.

[3] Aditya Joshi, Johan T. Halseth, Pentti Kanerva, "Language Recognition Using Random Indexing", Under Review As A Conference Paper At Iclr 2015, pp 1-7.

[4] Tanveer Ali, David Schramm, Marina Sokolova, Diana Inkpen, "Can I hear you? Sentiment Analysis on Medical Forums", International Joint Conference on Natural Language Processing, Nagoya, Japan, 14-18 October 2013, pp 667–673.
[5] Yihan Deng, Matthaeus Stoehr, "Retrieving Attitudes: Sentiment Analysis from Clinical Narratives", MedIR July 11, 2014, Gold Coast, Australia, pp 12-15
[6] Bo Pang, Lillian Lee, "Opinion Mining and Sentiment Analysis", Foundations and Trends in Information Retrieval , Vol. 2, Nos. 1–2 (2008) , pp 1–135, DOI: 10.1561/1500000001
[7] Despo Georgiou, Andrew MacFarlane , " Extracting Sentiment from Healthcare Survey Data: An evaluation of sentiment analysis tools ",Science and Information Conference 2015 July 28-30, 2015, pp – 1-10.
[8] Wingyan Chung , Tzu-Liang (Bill) Tseng ,"Discovering business intelligence from online product reviews: A rule-induction framework Mining opinion components from unstructured reviews: A review", Expert Systems with Applications 39 (2012) , pp 11870–11879
[9] Khairullah Khan , Baharum Baharudin , Aurnagzeb Khan , Ashraf Ullah ,"Ranked WordNet graph for Sentiment Polarity Classification in Twitter", Journal of King Saud University – Computer and Information Sciences (2014) 26, pp 258–275
[10] Arturo Montejo-Ráez, Eugenio Martínez-Cámara, M. Teresa Martín-Valdivia, L. Alfonso Ure˜na-López, "Sentiment Classification and Polarity Shifting", Computer Speech and Language 28 (2014) , pp 93–107
[11] Shoushan Li, Sophia Yat Mei Lee, Ying Chen, Chu-Ren Huang† Guodong Zhou, "Sentiment Classification and Polarity Shifting" , Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 635–643, Beijing, August 2010
[12] Jose, Sunny, and K Navaneetham. "A Factsheet on Women's Malnutrtion in India." Economic and Political Weekly. 43.33 (2008): 61-67. Web. 21 February 2013.
[13] Pathak, Praveen. "Economic Inequalities in Maternal Health Care: Prenatal Care and Skilled Birth Attendance in India, 1992-2006." PLoS ONE. 5.10 (2010): 1-17. Web. 7 February 2013. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2965095/>.
[14] Pednekar, Mangesh, Rajeev Gupta, et al. "Illiteracy, low educational status, and cardiovascular mortality in India." BMC Public Health. 11.567 (2011): 1-12. Web. 21 February 2013.
[15] Nayak, Madhabika, Vikram Patel, et al. "Partner alcohol use, violence and women." British Journal of Psychiatry. 196. (2010): 192-199. Web. 14 April 2013.
[16] Panda, Pradeep, and Bina Agarwal. "Marital violence, human development and women's property status in India." World Development. 33.5 (2005): 823-850. Web. 28 April 2013.