ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

Intelligent Data Mining of Social Media for Improving Health Care

Aradhana. S.Ghorpade

P.G. Student, Department of Computer Science& Engineering, DYPCET, Kolhapur, Maharashtra, India[1]

ABSTRACT: online Social media is providing massive opportunities for users to discuss their experiences and opinion with medicines and diseases, Pharmaceutical companies prefer social network monitoring within their IT departments, creating an opportunity for rapid dissemination and feedback of products and services to optimize and enhance delivery, increase turnover and profit and reduce costs. Online Social media can open the door for the health care sector to optimization service and patient care. The proposed system gives opinion related to medicine and most discussed side effects

KEYWORDS: Data Mining, Online Social Networks, neural networks, network based modeling.

I. INTRODUCTION

Intelligently extracting data from social network has attracted great interest in every field. The Health Informatics community has also used social media data extraction to simultaneously improve healthcare outcomes using consumer-generated sentiments.

With the massive growth of social media (i.e., reviews, forum discussions, blogs and social networks) on the Web, individuals and organizations are increasingly using public opinions in these media for their decision making. Potential customers also want to know the opinions of existing users before they use a service or purchase a product.

Healthcare providers could use patient opinion to improve their services. Physicians could collect feedback from other doctors and patients to improve their treatment recommendations and results. Patients could use other consumers' knowledge in making better-informed healthcare decisions. So the system is proposed to extract user's opinion related to a specific medicine. The networks are building to identify groups and influenced users. The most Discussed side effect related to that medicine is known in this system. Our paper is organized as follows: Firstly introduction. Next is literature Survey, the proposed system, experimental results and conclusion.

II. . LITERATURE SURVEY

Traditional social sciences use surveys and involve subjects in the data collection process, resulting in small sample sizes per study. So some of the researchers have used the following strategies to enhance the system

C. Corley, D. Cook, A. Mikler, and K. Singh, developed method of Text and structural data mining for influenza disease [1].Text and structural data mining of web and social media (WSM) provides a novel disease surveillance resource and can identify online communities for targeted public health communications (PHC) to assure wide dissemination of pertinent information. WSM that mention influenza are harvested over a 24-week period, 5 October 2008 to 21 March 2009. Link analysis reveals communities for targeted PHC. Text mining is shown to identify trends in flu posts that correlate to real-world influenza-like illness patient report data. It also brings to bear a graph-based data mining technique to detect anomalies among flu blogs connected by publisher type, links, and user-tags.

S. R. Das and M. Y. Chen proposed an algorithm for sentiment extraction for small talk on the web" [2]. Extracting sentiment from text is a hard semantic problem. They develop a methodology for extracting small investor sentiment from stock message boards. The algorithm comprises different classifier algorithms coupled together by a voting scheme. Accuracy levels are similar to widely used Bayes classifiers, but false positives are lower and sentiment

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

accuracy higher. Empirical applications evidence a relationship with stock values—tech-sector postings are related to stock index levels, and to volumes and volatility.

L. Getoor and C. Diehl [3], performed survey on link mining. In which they have discussed that many datasets of interest today are best described as a linked collection of interrelated objects. These may represent homogeneous networks, in which there is a single-object type and link type, or richer, heterogeneous networks, in which there may be multiple object and link types (and possibly other semantic information). Examples of heterogeneous networks include those in medical domains describing patients, diseases, treatments and contacts, or in bibliographic domains describing publications, authors, and venues. *Link mining* refers to data mining techniques that explicitly consider these links when building predictive or descriptive models of the linked data. This is an exciting, rapidly expanding area.

M. E. J. Newman, "Detecting community structure in networks [6] has reviewed algorithmic methods for finding common unities of densely connected vertices in network data.

A. Akay, A. Dragomir, and B. E. Erlandsson [4], developed a novel data mining method for extracting consumer opinion on diabetic disease based on the user post from the different forums.

Despite the extensive literature, none have identified influential users, and how forum relationships affect network.

III. PROPOSED WORK

The proposed system will intelligently mine data from online social media by using forum post and user feedback. Natural language processing and data mining techniques will be used for mining forum post and user feedback. At first using the Stanford nlp toolkit and self-organizing maps (SOMs) clustering exploratory analysis will be employed to assess correlations between user posts and positive or negative opinion on the medicine. Then network will be build using sub graph method .the module obtained will give the information broker users and most discussed side effect of medicines will be identified which can be used to improve care.

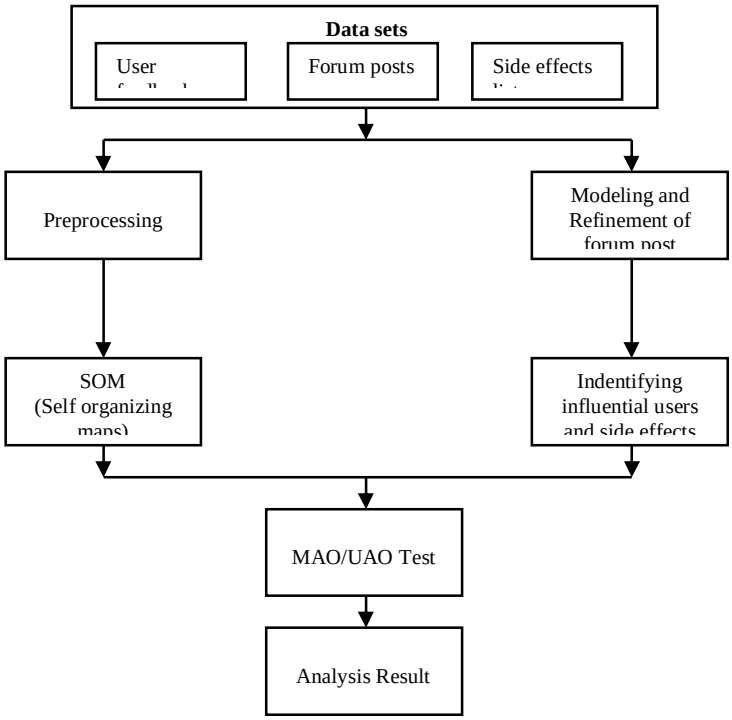


Fig: Proposed System Architecture

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

a) Data collection and preprocessing.

In Data collection process a single disease is selected and most discussed medicine related to that disease will be taken from various sites, forums and feedbacks related specific medicine. Processing will be performed on the raw text data collected using the language processing libraries and algorithms to look for the most common positive and negative words and their term-frequency-inverse document frequency (TF-IDF) scores within each post. The user opinion related to that medicine is obtained using nlk libraries.

b) Consumer sentiments analysis.

For this part of the analysis, all positive and negative words observed within the post and feedbacks are fed to the SOM to cluster the words and find the correlation between negative and positive words. Subgroups (neurons) were formed on the basis of their weights assigned in the previous module. The similar weighted words are group in same neuron and network based model is formed. This neuron shows the positive and negative words correlation.

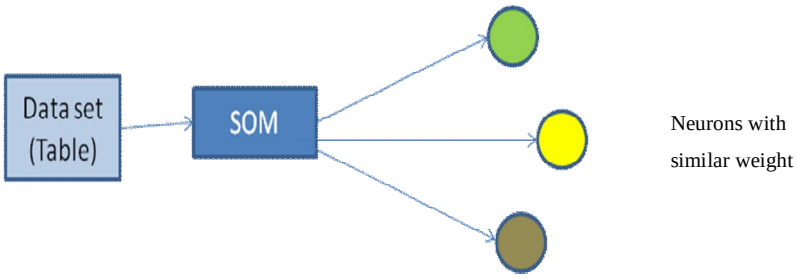


Fig: network based model using SOM.

c) Modeling forum posting:

The next step in our analysis is to build networks from forum posts and their replies. In a first step, we aimed at identifying *sub graphs* within our networks. After the sub graph were formed Influential users are users which broker most of the information transfer within network modules were identified whose opinion in terms of positive or negative sentiment towards the treatment is 'spread' to the other users within their containing modules. To obtain this the previously derived algorithm is used were [7] author proposed an approach in which transition probabilities for a random walk of length *t* (*t* being the Markov time) enable multi scale analysis.

d) Module average opinion and user average opinion:

In the second step of analysis, we devised a strategy for identifying influenced users .To this goal; we overlay the TF-IDF scores of the wordlist onto modules. The TFIDF scores within each module will thus directly used to calculate module average opinion (MOA) and user average opinion (UOA) with positive and negative words tf-idf score.

e) Identification of side effects and performance analysis.

On the basis of the MAO and UAO the influential users are find out and only that modules are considered for further analysis. The TFIDF scores within each module will thus directly reflect how frequent a certain side-effect is mentioned in module posts. .further the T -Test is applied to evaluate performance analysis of the obtained result .after the complete mining the most discussed side effects can be obtained.

IV. EXPERIMENTAL RESULTS

The experiments were conducted on the real word medicine review dataset which is collected from https://patient.info. Total 100 post were found on metformin type 2 medicine .this post are further used for processing .raw data is proposed and word with tf-idf score are obtained these words are assigned with weight .the positive and

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

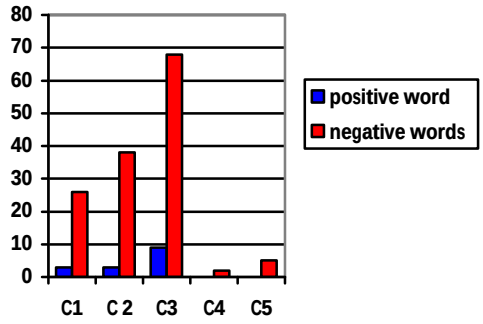
International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

negative opinion are obtained by using Stanford nlk libraries. Obtained words are fed for clustering. Similar weighed words and grouped in one cluster. This shows the correlation between positive and negative words.



.

Fig: chart of positive and negative word correlation using Som.

After the SOM analysis network is build by using post threads and replies and sub graph are identified. The side effect list is generated of the overall post collected.

Side Effects
cold
weakness
heartburn
diarrhea
tiredness
headaches
Irritability

Fig: side effect list.

From the modules obtained the most discussed side effects are found by comparing their overall tf-idf score to the list obtained within the module. The t-test is applied to analyze the final result according to the t-test performed weakness; diarreaha and tiredness are the most discussed side effects within the module.

Modules	Side effect	p-value
Module 1	weakness	P< 0.05
	Diarrhea	P< 0.01
	Tiredness	P< 0.01
Module 2	Diarrhea	P< 0.01
	Tiredness	P< 0.01

Table: Side-Effect Frequency and Location in Selected Modules

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor & UGC Approved Journal)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

V. CONCLUSION

We converted a forum focused on oncology into weighted vectors to measure consumer thoughts on the drug using positive and negative terms parallel another list containing the side effects. Social media can open the door for the health care sector in address cost reduction, product and service optimization, and patient care. These obtained results could be used as feedback loop for medicine manufacture companies. In future studies the post can be categorized on basis of their rankings or likes of post.

REFERENCES

[1] C. Corley, D. Cook, A. Mikler, and K. Singh, "Text and structural data mining of influenza mentions in web and social media," *Int. J. Environ.Res. Public Health, vol. 7, pp. 596–615, Feb. 2010.*

[2] S. R. Das and M. Y. Chen, "Yahoo! for Amazon: Sentiment extraction from small talk on the Web," *Manag. Sci., vol. 53, pp. 1375–1388, Sep.* 2007

[3] L. Getoor and C. Diehl, "Link mining: a survey," *SIGKDD Explore. Newsl.* vol. 7, pp. 3–12, Dec. 2005.

[4] Altug Akay (M'11), Network-Based Modeling and Intelligent Data Mining of Social Media for Improving Care.

[5] A. Akay, A. Dragomir, and B. E. Erlandsson, "A novel data-mining approach leveraging social media to monitor consumer opinion of sitagliptin," *J. Biomed Health Inform.* Vol: PP, Issue: 99.

[6] M. E. J. Newman, "Detecting community structure in networks," *Eur. Phys. J.*, vol. 38, pp. 321–330, Mar. 2004.

[7] E. Le Martelot and C. Hankin, "Multi-scale community detection using stability as optimization criterion in a greedy algorithm," *Proceedings of the 2011 International. Conf. erence on Knowledge Discovery and Information Retrieval (KDIR 2011), Paris, France: SciTePress, Oct. 2011,* pp. 216–225.