Intelligent Data Mining Technique of Social Media for Improving Health Care

Sopan Ganpat Sutar
Computer Science and Engineering Department
Government Engineering College, Aurangabad
Aurangabad (MH)-431001, India
sopan.sutar169@gmail.com

Abstract:- **A Data offers many facilities to the end users such as software, organization and platform go on. In this paper, we study about the wisely mining knowledge of social media. Social media becomes much popular from the health care information and Biomedical. This information is commonly shared so healthcare is improves and costs is decrease using opinion which is generated by user. We suggest investigation framework that give attentions on side effects of drugs and also focus on positive and negative response.**

To improve health care some Clinical documents are mostly useful because it's are free-text data sources. Clinical documents containing information related to symptoms and valuable medications. To extract a Data from large dataset it's become a very popular because users get various ideas from this filtered data. All Data Mining and Knowledge mining become popular because user are process on data and getting information of different area like health, Social, etc. After data processing we focus on users positive and negative opinions. We count this opinions and find out which medication is good, to decide this we also find out the side effects of the medications. Further we focus on the symptoms of the cancer patient. By taking the expert doctors suggestion, we list out the medication of the cancer according to the symptoms and we provide this medication or treatment to the user on our forum. We can expand our research into Data and Knowledge mining of social media and takes the users' views on various drugs of cancer. This daily updated data helps to pharmaceutical industry, doctors, hospitals, and medical staff, for effective future treatments.

Keywords: **Knowledge mining, Complex networks, social computing, Data mining, semantic web.**

I. INTRODUCTION

Social media is presenting countless opportunities for patients to give their opinion about particular drugs and devices, many organizations can also acquire feedback on their services and products [1]-[3]. Medical organizations are giving first priority to social network tracking in their Information Technology departments, creating possibility for speedy distribution and products related comments. It offers to beautify and optimize shipping. This optimization helps to decrease charges and finally it increases profit and turnover [4]. Facts of Social media are gathering for bio investigation also mentioned additionally [5]. Social media permits a VNE (virtual networking environment). We can extract Knowledge by using various computational tools. By using available networks making a group of social media is also a one of the way of knowledge extracting. Social network is just like a structure which is a collection of edges and nodes these nodes and edges are connected with each other in numerous relationships.

Due to the fast increase and improvement of digital textual content statistics made to be had in latest few years, text Mining, Data mining, statistic mining and mining of knowledge becomes more popular to convert such information in to beneficial records and meaningful knowledge.

Data mining can be considered as the extraction of raw or useless data from huge databases, many applications like healthcare systems, market analysis get advantages by such mined data and also they came to know how to extract useful data from a big amount of data. This extracted data is mostly useful to customer.

Users Internet struggling is come to be a completely famous with net chat the use of a social websites. Now humans will provide specific critiques on net website online, like a massive U. S. India range of people are using social web sites to tweet, chat. So this internet records become a totally critical element to peoples who want to get a few nice and negative statistics of very own field.

Many methods had been working like consisting of link mining [6], type via hyperlinks [7], predictions primarily based on items hyperlinks [8], lifestyles, estimation [9], object [10], institution, and subgroup detection [11], and data mining[12]. By using person's feedback Link prediction, biochemical marketing, online discussion agencies (and rankings) permit for improvement of answers.

Very Important Data are launch with discussion board method we have a completely treasured information related to fitness care because all subject matter going associated with most cancers and associated brought on and medicine. So we need to awareness in these valuable facts as assets. This all methods are to be had to get information and system on facts but this information having some limitation due to the fact this all able to paintings on pattern model simplest method we just technique on small amount of statistics and get a small pattern of facts .For preceding strategies there are one-of-a-kind assets from which statistics are collect and developer work in this information, Data source was government fitness monitoring, newspaper articles, on-line buying, and so forth. None of them diagnosed a social forum that impact on network dynamic and Data.

In the first degree of our modern have a look at, we hire exploratory evaluation the usage of our own dictionary

which contains positive and negative words. This dictionary is used to evaluate correlations between user posts and their response on drugs. This response is calculated in the form of positive and negative opinion on drugs. In the second o level we version the users and their posts on the basis of side effects and symptoms.

II. LITERATURE SURVEY

Many peoples are works on the cancer treatments to improve health care, Si Yan and Yanliang Qi[13] worked on cancer research they used the 3 various text mining tools namely MedLEE , HITEx and caTIES. These tools are developed by Columbia University and Harvard University these tools helps to extract medical information and diagnosis from pathology report. After extracting information they made one hypothesis which contains the specific type of cancer and drugs used for same.

T. Anisha and Mr N.[14] Thulasi measures users views on drug Erlotinib using positive and negative views along with their side effects. This method helps to find out the users response of the drug Erlotinib on lung cancer. They used the self- organizing map and find out the frequently word from the users' tweets.

Akay A., Member IEEE, proposed a Novel Data Mining method to find out the experience of drug Januvia which is used for *diabetes* [15]. In this method he took the opinion of the users on drug Januvia and find out the positive and negative results on it. So by using this users opinion's it helps to provide best solutions for public health.

Sentiment has an idiosyncratic component and aggregation of sentiment across stocks tracks index returns extra strongly than with person shares. Preliminary proof suggests that market hobby impacts small investor sentiment. Thus, the algorithms advanced on this paper can be used to assess the effect on investor opinion of control announcements, press releases, information, and regulatory changes. Most facts retrieval systems use prevent word lists and stemming algorithms. However, we've got located that spotting singular and plural nouns, verb bureaucracy, negation, and the prepositions can produce dramatically one of the kind text type consequences.

III. PROBLEM STATEMENTS

1. Old system considers government health monitoring, newspaper articles, online shopping, etc.
2. Only focus on Negative positive tweet.
3. Just work on Simple Dataset.
4. Result will be complicated
5. Not properly use Text mining to find Symptoms and Medications

IV. PROPOSED SYSTEM

1. New Cancer forum that including cancer related information and data flow.
2. New Approach to use Forum on Data processing which provide valuable knowledge.
3. Get cancer related Data Mining on different opinion to find out Data and important assets.
4. Getting knowledge base information related cancer.
5. Get Positive and Negative opinion related medicines.
6. Provide social health related important information to patient.
7. Find out the treatment and its side effect to user and provide a solution for this.
8. Find Symptoms and Medication which is very useful to everyone.

In this Paper we just added a new module of symptoms and medication related symptoms. We not providing a positive and negative against a treatment but we also providing which medication are used for particular symptoms.
We providing a data i.e. if user get side effect related any treatment then we providing what we can get solution if user has side effect.

We also add health care as well as social care, environmental care from data and knowledge mining. User can add own opinion on forum also.

V. ARCHITECTURE

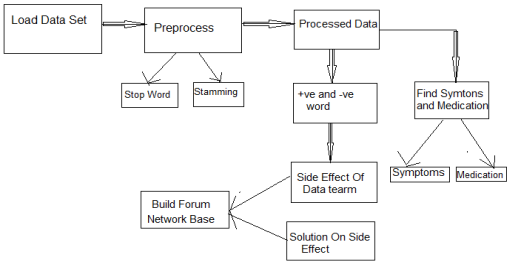

Fig.1 System architecture

VI METHODS

A. *Get Dataset*

We are taking data from cancer forum site. This data is in the form of users tweets related to all cancer types and its treatments. Also this module provides the facility to live tweet and these tweets are taken as an input dataset for processing. These data is related to treatments and experience of particular drugs on a cancer so it is a raw data from this we have to find out the how many peoples give positive response. After collecting this dataset, data will send to next process i.e. Text processing.

B. *Text Processing*
Text processing is a process in which we remove the stop words and text stemming
1) Stop words removal: The words rather than natural language words are the stop words. In short stop words are words which are meaningless.
2) Text stemming: Inflected and derived words are removed in stemming process. These are removing on their stem base or root form. It generally a written word forms.

C. *Pattern taxonomy process*
When the imported document having big size so its' become difficult to the text processor to process dataset. So to overcome this problem the document or dataset is break in to the small paragraphs. Now each paragraph is considered as separate document. Some terms are extracting from positive document and these terms are extracting in to each document.

D. *Find out positive and negative words*
Processed data is useful to find out positive and negative comments of user. To find this we used our own predefined dictionary in which we add the positive and negative words. To avoid the repeated words we used TF-IDF algorithm. Term frequency (TF) count the frequency of words found in the document it means it how many times occurs a particular word.
IDF (Inverse document frequency) calculate the percentage of term occurs in main document.

E. *Calculate satisfaction or dissatisfaction rate of user*
After find out negative and positive response of each user we can calculate the satisfaction and dissatisfaction ratio of users so that we conclude that how many users are given positive response. After knowing which users are satisfy with cancer treatment we can study their tweets and find out which drugs they used for cancer.
SFR (Satisfaction Ratio) and DSFR (Dissatisfaction Ratio) can calculate by following formula

$$SFR = \frac{Sum\ (PW) \times 100}{Sum\ (all)}$$

$$DSFR = \frac{Sum\ (NW) \times 100}{Sum\ (all)}$$

Where,

Sum (PW) =All positive words
Sum (NW) =All negative words
Sum (all) = PW+NW

SFR and DSFR are helps to know how many users are get benefit of our forum.

F. *Calculate Symptoms and Medication*
By using Our predefined dictionary of symptoms words we can find out the symptoms from users tweets. By taking expert doctor's opinion or using online Medical dictionary we can list out the medication according to symptoms. We can add this information in our database for best result. So that the user can select their symptoms and came to know medication for cancer so that the health care is improve by using our forum because its update regularly by users and admin. Experimental results show that multi-view NMF is a preferable method for clinical document clustering. Moreover, we find that using extracted medication/symptom names to cluster clinical documents outperforms just using words.

VII. RESULTS

With using this technique we develop a new cancer forum which is totally helping a patient of cancer. We process on network data and find out positive and negative data from dataset. Positive data focus on the data which are necessary to treatment and negative data focus on a side effect of treatment.

With a text mining we process on data find out filter data and getting all cancer related information. Also we can use methods to filter data and find out Symptoms list and Medication list depend on dataset. So user or patient can easily select symptoms and find out medication related to symptoms.

FINAL POSTANALYSIS WORDLIST

Positive +	Negative -
Agree	Bad
Benefit	Concerns
Comfort	Damage
Comfortable	Dangerous
Effective	Died
Enjoy	Difficult
Favorable	Discomfort
Good	Error
Grateful	Failure
Great	Fear
Helped	Impossible
Helpful	Isn
Helping	Lack

FINAL SIDE EFFECTS WORDLIST

Acne
Cachexia
Headaches
Itching
Lesion
Pneumonia
Rash
Tremor
Weakness
Vomiting

Fig. 2. Positive and negative word list with side effects from old method

From the data Processing we just find out a positive and negative word. This is come from the cancer forum or user tweets. Also we find out the teams that create a side effect for patient treatment. Positive and negative word having a very effect on data filtration because data related forum its very broad so we focus on the important terms that direct positive and negative effect on cancer related activity.

User Name	PW	NW	SFR	DSFR
Sopan	5	3	62%	38%
Ravi	10	2	83%	17%
Sonali	17	7	70%	30%
Amar	4	6	40%	60%
Sachin	3	9	27%	73%

Table1- SFR and DSFR of user from their tweets.

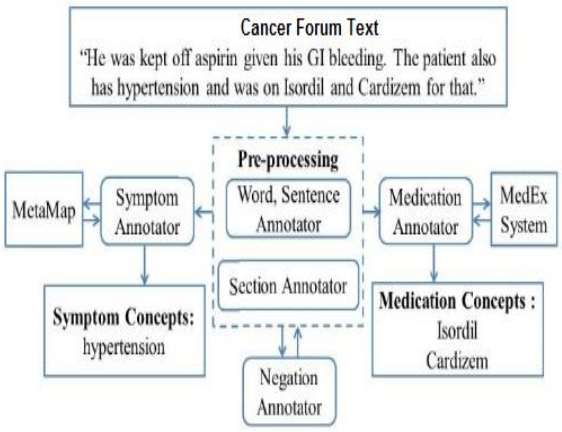

Fig. 3.To find Symptoms and Medication from Cancer Forum Text Data by proposed method

With using forum Data we are process on data and find out Suspect statement where we can process data. After getting users positive and negative views we find out the medication and symptoms from positive users tweets and add this information in our forums database.

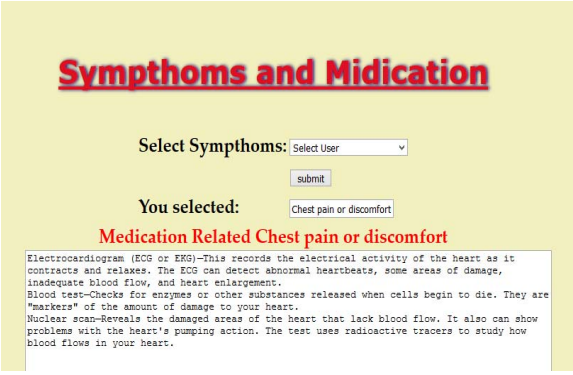

Fig. 4. Medication according to symptoms

Above figure 4 shows the medication for the particular symptoms which is tweet by users. In old system we can only find out the positive and negative side effect of the particular drugs used by patient. But in the proposed technique we again process the data and find out the users view on the drugs in the form of symptoms and finally we compare all results and find out which medication is good for particular diseases on the basis of symptoms and side effect.

VII CONCLUSION

In this paper, we build an integrating system to extract treatment, side effect symptom/medication names from unstructured/semi-structured Data from cancer forum. The overall system conations patient treatment, positive Negative effect, Medication name, Symptoms detail, many data mining techniques [18] have been proposed in the last decade. These techniques include association rule mining, frequent text mining [19], sequential text mining, maximum pattern mining. However, using these discovered knowledge (Cancer Forum) in the field of text mining is difficult and ineffective. As size of information present on the internet has taken a shape of the giant it has become a necessity to increase the efficiency of the search Data. Data and knowledge mining on data is very important because we are getting a valuable information which is not easily available, and all information are real time information.

In our future work, we may consider using other information, such as patient's age/gender/demographical information, to improve clustering performance; and also explore intrinsic relationships among different views. We also plan to use the document clustering results to improve medication recommendation as discussed in our former work.

REFERENCES

[1] Ochoa, A. Hernandez, L. Cruz, J. Ponce, F. Montes, L. Li, and L. Janacek. "Artificial societies and social simulation using ant colony, particle swarm optimization and cultural algorithms," New Achievements in Evolutionary Computation, P. Korosec, Ed. Rijeka, Croatia: InTech, pp. 267–297, 2010.

[2] W. Cornell and W. Cornell. (2013). How Data Mining Drives Pharma: Information as a Raw Material and Product. [Online]. Available: http://acswebinars.org/big-data

[3] L. Toldo, "Text mining fundamentals for business analytics," presented at the 11th Annual Text and Social Analytics Summit, Boston, MA, USA, 2013.

[4] L. Dunbrack, "Pharma 2.0 – socialmedia and pharmaceutical sales and marketing," in Proc. Health Ind. Insights, 2010, p. 7.

[5] 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.

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

[7] Q. Lu. and L. Getoor, "Link-based classification," in Proc. 20th Int. Conf. Mach. Learning, Washington, D.C., USA, 2003, pp. 496–503.

[8] Ng, A. Zheng, and M. Jordan, "Stable algorithms for link analysis," in Proc. SIGIR Conf. Inform. Retrieval., New Orleans, LouisianaLO, USA, 2001, pp. 258–266.

[9] Z. Lacroix, H. Murthy, F. Naumann, and L. Raschid, "Links and paths through life sciences data sources," in Proc. 1st Int.Workshop Data Integr. Life Sci., Leipzig, Germany, 2004, pp. 203–211.

[10] J. Noessner, M. Niepert, C. Meilicke, and H. Stuckenschmidt, "Leveraging terminological structure for object reconciliation," in The Semantic Web:Research and Applications, Heidelberg, Berlin: Springer, 2010, pp. 334–348

[11] J. Huan and J. Prins, "Efficient mining of frequent subgraphs in the presence of isomorphism," in Proc. 3rd IEEE Int. Conf. Data Mining, Melbourne, Florida.FL, USA, 2003, pp. 549–552.

[12] D. Hand, "Principles of data mining," Drug Safety, vol. 30, pp. 621–622, Jul. 2007

[13] Si Yan and Yanliang Qi, "Apply Text Mining to Advance Cancer Research", IJPMBS, Vol. 4, No. 2, April 2015.

[14] T.Anisha1, Mr.N.Thulasi2," Improving Health Care Based on opinion Mining Using KNN",T.Anisha1, Mr.N.Thulasi2,IJIRSET, Vol. 5, Issue 4, April 2016.

[15] A.Akay and Bjorn-erik Erlandsoon,"A novel data mining platform leveraging social media to monitor outcomes of Januvia". In IEEE EMBS Osaka,Japan,3-7 july,2013.

[16] Taskar, M. Wong, P. Abbeel, and D. Koller, "Link prediction in relational data," in Proc. Adv. Neural Inform. Process. Syst., Vancouver, B.C. Canada, 2003.

[17] D. Liben-Nowell and J. M. Kleinberg, "The link prediction problem for social networks," J. Am. Soc. Inform. Sci. Technol., vol. 57, pp. 556–559, May 2007.

[18] J. Hans and M. Kamber, Data Mining: Concepts and Techniques. 2nd ed.Burlington, MassMA, USA: Morgan Kaufmann, 2006.

[19] 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.201