

# Extracting Knowledge from Social Media to Improve Health Informatics

Mr. V.R.Nagarajan<sup>1</sup>, Monisha.P.M<sup>2</sup>

Assistant Professor, Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu<sup>1</sup>

M. Phil. Scholar, Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu<sup>2</sup>

**Abstract:** Social media provide an intelligence information extraction for the area of Health Informatics, including Bioinformatics, Image Informatics, Clinical Informatics, Public Health Informatics, etc. To improve the health care outcomes with reduced cost we propose the framework which concentrates on positive and negative symptoms of the disease and also the side effects of the treatment. For this health care we take most prevalent cancers: lung cancer. Our proposed approach we employ investigatory analysis using the SSOMs to check the correlations between user posts and positive or negative judgment on drug. Then we use the hierarchical clustering, it measures the starting with an empty network of  $n$  vertices and no edges, one add edges between pairs of vertices in order of decreasing similarity, starting with the pair with strongest similarity.

**Keywords:** Social media, Health Informatics, lung cancer, SOM and hierarchical clustering.

## I. INTRODUCTION

Social media is one to extracting the information from internet. Now a day it is used for extracting the data of patient's to know the experiences of patient symptoms. Social media, assorting from individual messaging to live foras, is providing immeasurable opportunities for patients to converse their experiences with drugs and devices. Social media allows message, contribution, gathering information, and distribution in the healthcare space. Healthcare space is one which contains the information of the patients with their permissions. It provides an effective social networking environment. The proper way to mine information and drift from the knowledge is cloud. Using network based analysis methods it model the social media such as Facebook, Twitter, WebMD, etc [1]. These analyses are made of nodes which may be individual or organization that connect the network in various associations such as interests, friendship, kinship, etc. Hence the representation of the information would be a graphical, which is convenient for the user visualization. But in traditional level it uses the surveys to collect the data from the patients and which were resulting in small sample sizes per surveys and studies. With the help of social media, they were additional content is willingly available, particularly when shared with web-crawling and scraping software it allow real-time monitoring of modifications inside the network. In social network it provides an in-depth knowledge of social network dynamics. A network model could be used for replication studies of different network aspects such as mining how the users broadcast knowledge among themselves (news about epidemic or drugs' difficult effects). Other example is analyzing the augmentation of certain edges of networks and how certain information involves the enhancements (e.g., how certain user communities develop based on ordinary benefit about explicit diseases). In social media the duty of social networks involves data collection in

difficult manner. For this it used several techniques to employee, such as links, existence, estimation, object, group, and subgroup detection, and mining the data, link mining, classification through links, predictions based on objects. Other techniques are also based on the user feedback they were Link prediction, viral marketing, online discussion groups (and rankings) allow for the development of solutions based on user feedback [2].

In social network, the data are heterogeneous, multi-relational, and semi-supervised, to gather these data is difficult. One method of extraction is link mining or relationship extraction, which provide the effective way for combining the social network using hypertext and web mining to extract data from web. Combining this, it involves classification; prediction, estimation, reconciliation, detection and finally extraction were done to extract the patient information. There are two simple ways to mine this information: 1) crawling and 2) scraping, with crawling it afford site Application Programming Interfaces (APIs) then second rendered information from HTML pages [3].

In our approach we employ investigatory analysis using the self-organizing maps (SOMs) to check the correlations between user posts and positive or negative judgment on the drug. Then we use the hierarchical clustering, it measures the starting with an empty network of  $n$  vertices and no edges, one add edges between pairs of vertices in order of decreasing resemblance, initial with the pair with strongest resemblance. Then we form the users and their posts using a network-based hierarchical clustering approach. These approaches were used for identifying user communities (modules) and influential users therein. Our approach successfully searches for potential levels of organization (scales) within the networks and uncovers dense modules. This enables us to find the possible partition using complete linkage. We consequently

improve the rescue modules with word regularity information from module-contained user's posts to obtain local and global measures of user's opinion and raise flag on potential side effects of Erlotinib, a drug used in the treatment of one of the most prevalent cancers: lung cancer. It provides a better result when compared to the previous work [4, 5].

## II. LITERATURE REVIEW

From author Jun Huan, et al [6] frequent sub graph mining is an active research topic in the data mining community. A graph is a common form to signify data and has been used in many domains like informatics and bioinformatics. Extracting patterns from graph databases is challenging since graph related operations, such as sub graph testing, normally have elevated time complexity than the corresponding operations on item sets, sequences, and trees, which have been considered broadly. From this paper, they propose a novel frequent sub graph mining algorithm: FFSM, which employs a vertical investigate method within an algebraic graphical structure we have developed to reduce the number of superfluous candidates proposed. Our experimental study on artificial and real datasets demonstrates that FFSM achieves a substantial performance gain over the current start-of-the-art sub graph mining algorithm gSpan.

From Sanjiv R. Das et al [7] we develop a methodology for extracting small investor sentiment from stock message boards. Five different classifier algorithms attached by a determination scheme are found to execute well alongside human and statistical standard. Time series and cross-sectional aggregation of message information progress the superiority of the resultant sentiment directory. Empirical applications evidence a relationship with stock returns – visually, using phase-lag analysis, pattern recognition and arithmetic methods. Sentiment has an idiosyncratic component and aggregation of sentiment across stocks tracks index revisit more strongly than with entity stocks. Preliminary evidence recommends that market activity influences small investor sentiment. This is used to assess the impact on investor estimation of management proclamation, press liberate, third-party news, and authoritarian changes.

From Noémie Elhadad, et al [8], Social media sites are a major source for non-curated, user-generated feedback on virtually all products and services. Users increasingly rely on social media to disclose sometimes serious real-life incidents rather than visiting social communication channels. This valuable, actionable, user-generated information, if extracted reliably and robustly from the social media sites, has the potential to have a positive impact on critical applications related to public health and safety, and beyond. Unfortunately, the extraction and presentation of actionable information from social media where the output of the extraction process is used to take concrete actions in the real world are not well supported by existing technology. Traditional information extraction approaches do not work well over the highly informal,

noisy, and ungrammatical text common in social media, and they do not handle the extraction and aggregation of the rare content that important applications need to extract from high-volume streaming sources. In our ongoing collaborative project between Columbia University and the New York City Department of Health and Mental Hygiene (DOHMH), we aim to address these gaps in research and technology for one important public health application, namely, detecting and acting on food borne outbreaks in New York City restaurants. Thus far, we have been able to address these issues successfully and have used one social media site to identify and follow up on several food borne outbreaks that had not been reported through conventional channels.

From Erwan Le Martelot et al [9], nowadays, networks are roughly everywhere. The community detection received an increasing attention as a way to uncover the formation of networks by grouping nodes into communities more densely connected internally than superficially. Yet most of the effective methods available do not consider the possible levels of organization, or scales, a network may encompass and are therefore limited. In this paper we present a method compatible with global and local criteria that enables fast multi-scale community finding. The method is derived in two algorithms, one for each type of criterion, and executed with 6 known criterion. Discovery communities at various scale is a computationally luxurious task. Therefore this work puts a strong emphasis on the reduction of computational complexity. Several heuristics are commenced for speed-up purposes. Experiments exhibit the efficiency and exact of our method with respect to each algorithm and criterion by testing them against large generated multi-scale networks. This work also offers a assessment between criteria and between the global and local approaches.

From Matthew Herland, et al [10], The amount of data produced within Health Informatics has grown to be quite vast, and examination of this Big Data awards potentially immeasurable possibilities for information to be gained. In addition, this information can progress the quality of healthcare accessible to patients. However, there are a number of problems that arise when dealing with these vast number of data, especially how to analyze this data in a reliable manner. The basic objective of Health Informatics is to take in real world medical data from all levels of human existence to help proceeds our understanding of medicine and medical practice. This paper will shows topical research using Big Data tools and move toward for the analysis of Health Informatics data gathered at numerous levels, counting the molecular, tissue, patient, and inhabitants levels. In addition to gathering data at multiple levels, multiple levels of questions are addressed: human-scale biology, clinical-scale, and epidemic-scale. We will also analyze and examine possible future work for each of these areas, as well as how coalesce data from each level may provide the most promising approach to gain the most knowledge in Health Informatics.

From Erwan Le Martelot et al [11], many real systems can be represented as networks whose analysis can be very informative regarding the original system's organization. In the past decade community detection received a lot of attention and is now a very active field of research. Recently stability was introduced as a new measure for partition quality. This work investigates stability as an optimization criterion that exploits a Markov process view of networks to enable multi-scale community detection. Several heuristics and variations of an algorithm optimizing stability are presented as well as an application to overlapping communities. Experiments show that the method enables accurate multi-scale network analysis.

From Altug Akay et al [12] a novel data mining method was developed to gauge the experience of the drug Sitagliptin (trade name Januvia) by patients with diabetes mellitus type 2. To this goal, we devised a two-step analysis framework. Initial exploratory analysis using self-organizing maps was performed to determine structures based on user opinions among the forum posts. The results were a compilation of user's clusters and their correlated (positive or negative) opinion of the drug. Subsequent modeling using network analysis methods was used to determine influential users among the forum members. These findings can open new avenues of research into rapid data collection, feedback, and analysis that can enable improved outcomes and solutions for public health and important feedback for the manufacturer.

### III. PROPOSED SYSTEM-SSOM

Our proposed model includes some advance detection technique of inter-social dynamics, which incorporates the ratings, relationships and rankings. This also provides the word snippet option for formal language input in the drug review forums. This helps to predict the list of terms related to the medical domain which are drug names and disease with their symptoms. Mapping consumer opinion on drugs, treatments and other medical services is the major aim of our current proposal. The SSOM mapping framework has the ability to identify the contextual posting on the medical forums.

SSOM aims to discover knowledge from huge unstructured text in the drug review forum. The reviews were first transformed to as set of feature vectors in which every component corresponds to a different word. And the weight for every component has been calculated with the use of term frequency identification method. These vectors and its values are taken as the input, which is used to train the SSOM. In SSOM two maps are proposed which are named as word mapping and document mapping, which helps to label the neurons in the map. The SSOM results the reviews with the above two steps along with inter social dynamic maps.

#### a. Data Search and gathering

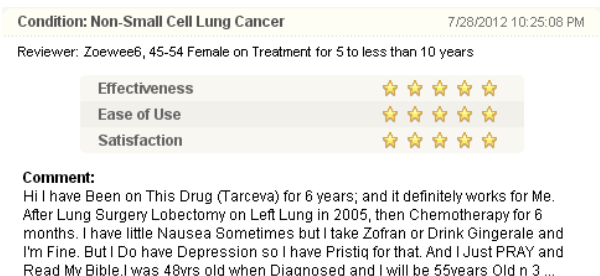
Data search and gathering is used to collect the relevant data for our research. Collection of data helps to provide a report about the patient details. For this our

method used for gathering the dataset of the drugs, for this we searched the most popular cancer message boards and after collecting data then focused on the number of posts on lung cancer. We concentrate on lung cancer this is because most of the recent statistics in common diagnosed cancer in both the sexes. After this we prepared the list of drugs used for the lungs cancer patients.

The drug review dataset has been collected from the WEBMD domain. In WEBMD domain the following drug based reviews are extracted for the experiments.

DRUG NAME	USER REVIEWS
cisplatin intravenous	28 User Reviews
doxorubicin intravenous	4 User Reviews
mechlorethamine injection	2 User Reviews

Fig1 (a): drug review dataset



Condition: Non-Small Cell Lung Cancer 7/28/2012 10:25:08 PM  
Reviewer: Zoewee6, 45-54 Female on Treatment for 5 to less than 10 years

Effectiveness	★★★★★
Ease of Use	★★★★★
Satisfaction	★★★★★

**Comment:**  
Hi I have Been on This Drug (Tarceva) for 6 years; and it definitely works for Me. After Lung Surgery Lobectomy on Left Lung in 2005, then Chemotherapy for 6 months. I have little Nausea Sometimes but I take Zofran or Drink Gingerale and I'm Fine. But I Do have Depression so I have Pristiq for that. And I Just PRAY and Read My Bible. I was 48yrs old when Diagnosed and I will be 55years Old n 3 ...

Fig 1 (b): sample drug review with ratings

The most common used drug is Erlotinib in the message board. We collect this from many Forums and also used the web from social media and form the cluster of the data using the hierarchical clustering for similarity measures. The idea behind this technique is to develop a measure of similarity  $x_{ij}$  between pairs (i, j) of vertices, based on the network structure one is given. In hierarchical clustering method it divides the network into smaller quantity and smaller communities. At the beginning of the algorithm it present n components consisting of a single vertex each, and at the end there is just one component containing all vertices. Hierarchical clustering performed using this measure should add edges to the network in order of increasing  $x_{ij}$ , not decreasing.

#### b. Preprocessing Data

Preprocessing is the method of preparing the data and eliminates the missing data. It helps to avoid the error in the dataset. Adding the missing data may occur inconvenient to the selection process. So there may be confusion occurs, so to avoid this error we use the preprocessing method. Here we use preprocessing techniques to check and removes the stop words. The techniques involve tokenizing, stop word filtering, stemming and eliminating auxiliary information's. The final component which is a perfect Processed Data will be taken as an input in SSOM. This helps to look for the most

common positive and negative words, and their term-frequency-inverse document frequency [13].

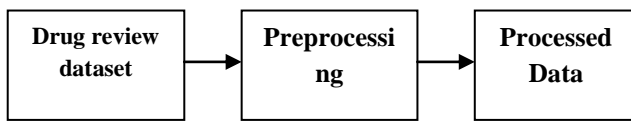


Fig2: Data Preprocessing

For the preprocessing the stage we upload the data into the readable document, then uploaded documents was processed in the second stage, here the documents were processed into data, here it filter the data and remove the noise data, after that it measures the set of variables, finally processed data were contains a specific term frequency.

**Tf-idf weighting:**

We combined the definitions of term frequency and inverse review frequency, to produce a composite weight for each term in each review. The tf-idf weighting scheme assigns to term  $t_d$  a weight in document  $d$  given by

$$tf-idf_{t,d} = tf_{t,d} \times idf_t$$

In other words,  $tf-idf_{t,d}$  assigns to term  $t_d$  a weight in document  $d$  that is

1. Highest when  $t_d$  occurs many times within a small number of documents.
2. Lower when the term occurs fewer times in a document or occurs in many documents.
3. Lowest when the term occurs in virtually all documents.

At this point, this may view each review as a vector with one component corresponding to each term in the lexicon, along with a value for each component.

**c. Categorizing data**

Categorizing the data is used to list the data in the relevant list. After preprocessing stage, the data in the database were partitioning into the relevant field, so only it will know the effects of the patients [14]. Based on the patient drugs and their category it will segment the data into its relevant fields. Based on our approach it reflected the negativity of a negative word and the positivity of a positive word in context. For this categorizing the data we use Centroid-Base Classifier to find out the aspects of the patients. It works when a new instance is ready to be classified it is assigned to the class consisting of the training samples whose average centroid is closest to the new instance. To find a patient can be classified as either low or high risk. If the data contains any keywords such as “not”, “no”, then the system performs re-classification with respective adjacent word.

**d. Creating connectivity using SSOM**

SSOMs are neural networks that produce low-dimensional representation of high-dimensional data. SSOM (Semantic

Self-Organizing Maps) to check the correlations between user posts and positive or negative judgment on the drug.

**New\_Node:** Create a new SSOM network

**Training Process:** Apply SSOM training algorithm on input document vectors.

**Mapping Process:** classifies the new input vector based on the training.

This approach was used for results validation. Within the network, a layer signifies the output space with every neuron dispense a specific weight. The weight values replicate on the cluster content. The SSOM shows the data to the network, fetch together similar data weights to similar neurons [15].

**e. Sub-graph and Side Effects Identification**

This method is used to convert the forum posts into the directional network, we create a path init to indentify the relevant group of the similarity measures, and this contains the internal and external units of the data. And scale it into local and global criteria for identifying the modules, to maximizing the measuring partition quantity which is stability to represent the nodes and edges which where the possible transitions in the unit. After that we find the side effects of the drugs and provide a result [16, 17].

**IV. EXPERIMENTAL RESULTS**

The experiments were conducted on the real word drug review dataset which is collected from WEBMD and the URL has specified below. <http://www.webmd.com/drugs/drugreview>. There are 1007 reviews in the website and those dataset has been experimented in SSOM. From the experiments, the system detects the best drug finding based on text reviews and ratings. Our system finds best drug from a set of review which contains 3 drugs named as Erlotinib, Cisplatin intravenous and doxorubicin. From the textual review, contextual information ratings and ratings the system found the Erlotinib drug has more positive review than others.

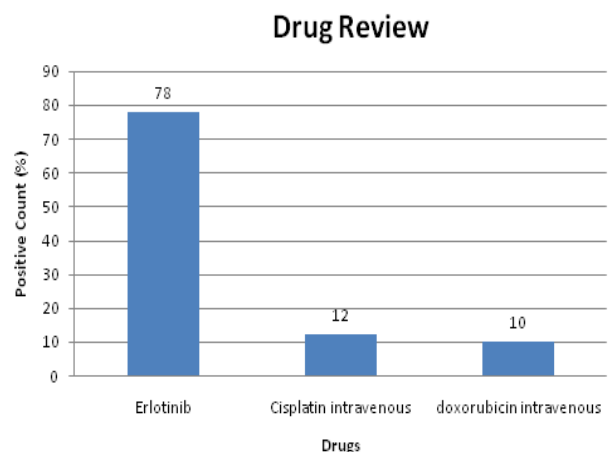


Fig.3: Drug comparison chart

There are two main parameters to evaluate the performance of the proposed approach SSOM, which are clustering quality which is measure with error rate and computation time.

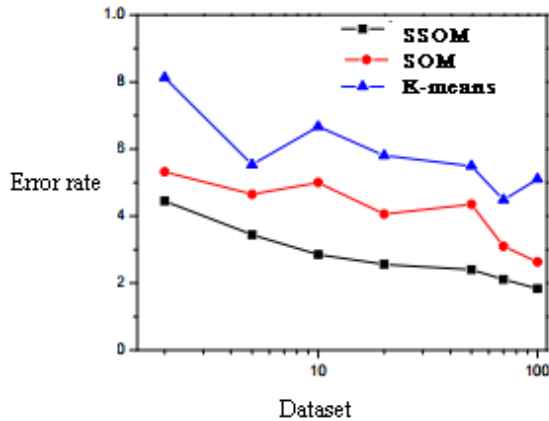


Fig 4: Error rate comparison chart with existing clustering techniques.

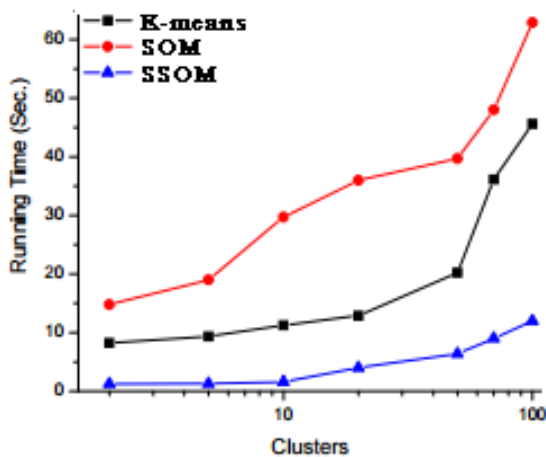


Fig 5: Time comparison chart with existing techniques.

From the above comparison charts, SSOM performance has been analyzed. The comparison has been studied with the existing document clustering techniques such as SOM and K-means.

#### IV. CONCLUSION

Social media is one of the emerging fields which provide a data on web for accessing the data through web. Recently it was used for the field of Health Informatics. Research done in Health Informatics to find the symptoms and side effects of the drugs based on the data mining approach. For this we use the approach called SOMs and hierarchical clustering. This paper shows the framework which concentrates on positive and negative symptoms of the disease and also the side effects of the treatment in prevalent cancers lung cancer.

#### REFERENCES

- [1] Altug Akay, "A Novel Data-Mining Approach Leveraging Social Media to Monitor Consumer Opinion of Sitagliptin".
- [2] Altug Akay, "Network-Based Modeling and Intelligent Data Mining of Social Media for Improving Care".
- [3] M. E. J. Newman, "Detecting community structure in networks".
- [4] D. Hand, "Principles of data mining," *Drug Safety*, vol. 30, pp. 621–622, Jul. 2007.
- [5] L. Dunbrack, "Pharma 2.0 – social media and pharmaceutical sales and marketing," in *Proc. Health Ind. Insights*, 2010, p. 7.
- [6] Jun Huan, Wei Wang, Jan Prins, "Efficient Mining of Frequent Subgraph in the Presence of Isomorphism".
- [7] Sanjiv R. Das, Mike Y. Chen. "Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web".
- [8] Noémie Elhadad, et al "Information Extraction from Social Media for Public Health".
- [9] Erwan Le Martelot, "Fast Multi-Scale Detection of Relevant Communities".
- [10] Matthew Herland, et al, "A review of data mining using big data in health informatics".
- [11] Erwan Le Martelot, "Multi-Scale Community Detection using Stability Optimisation".
- [12] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*. New York, NY, USA: Cambridge University Press, 1994, pp. 825
- [13] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E*, vol. 69, p. 026113, 2004.
- [14] M. E. J. Newman, "Scientific collaboration networks: II. Shortest paths, weighted networks, and centrality." *Phys. Rev. E* 64, 016132 (2001).
- [15] J. Vesanto, J. Himberg, E. Alhoniemi, and J. Parhankangas, "Self-Organizing Map in MATLAB: The SOM Toolbox," in *Proc. Matlab DSP Conf.*, Espoo, Finland, 1999, pp. 35–40.
- [16] A. Anna, C. Liza, A. L. C. Agero, S.W. Dusza, C. Benvenuto-Andrade, K.
- [17] J. Busam, P. Myskowski, and A. C. Halpem, "Dermatologic side effects associated with the epidermal growth factor receptor inhibitors," *J. Am. Acad. Dermatology*, vol. 55, pp. 657–670, 2006.