ISSN 2229-5518

Efficient Detection Technique and Algorithm for Health Care Improvement in Social Networks

E.SOWMYA, R.MANIKANDAN

Abstract— Recently, the large scale use of expertise has a significant impact on the worldwide health care sector. In social network, the number of users is spending their time for sharing communication and distributes content simply along with other users. When social networking sites distribute the essential reason of online communication and message, detailed objectives and outline of usage differ extensively across diverse services. By using data mining techniques the most important knowledge is extracted from the social network which is attracted tremendously in the field of biomedical and health informatics community. The several methods are suggested such as link prediction, hierarchical clustering and subgraph detection techniques to handle the health care data in social networks. However, the preceding research methods are issue with long time computation, inaccurate results and inefficiency. To overcome the above mentioned problems, we enhance a framework to analyze the forum reviews more effectively. This framework is focused on the analysis sentiments of positive, negative, neutral, side effects of treatment and recognizes posts, likes and comments of user's forum. The proposed work is identifying the user communities and influential users based on their opinions of cancer treatment. We introduced a self organizing map (SOP) for analyzing the word frequency from the user's forum posts, likes and comments. This research work is used network based modelling method to model the user's forum communications and improved the stability quality measure. It is used to discover the consumer opinions and recognize influential users within the retrieved modules by using information derived from both word-frequency data and network-based properties. The experimental results concluded that the proposed system is better than the existing system in terms of effective performances.

Index Terms— : Big data, Data mining, Semantic web, Social network, Opinion mining.

1 INTRODUCTION

Big data analytics is a development area along with the prospective to produce useful insight in healthcare segment. A collection of massive and complex datasets are complicated to process by using general database management tools. The healthcare industry traditionally has caused huge amounts of data, driven through record keeping, observance and authoritarian requirements, and patient care [1]. Data mining techniques also can be used in large big data applications to provide fast and accurate results. In medicinal science there is great scope for data mining applications. Analysis of health care, patient profiling and history creation are the few examples. Mammography is the technique used in breast cancer discovery. Radiologists face lot of complexities in tumors recognition and CAM (Computer Aided Methods) could assists to the medical staff.

In social network, the large volume of information is presented and extracted those information using data mining techniques. The social media provides various usages in several fields such as health care, fraud detection, medical science and banking. In the application of biomedical and healthcare, mining information from social network is producing number of advantages about the technology, updates of present information in the medical field [2]. Social media is providing certain opportunities for patients to discuss their experiences with drugs and devices. Social media activates a virtual networking environment and modeling social media using available network modeling as well as computational tool is one way of mining knowledge.Social networks allow individuals to switch over information on behalf of themselves or of others on such subjects such as the knowledge of bodily symptoms, medical analysis and treatment options, difficult treatment effects, sources of remedial evidence, familiarity along with individual providers and sentiments about their quality. This knowledge is also shared more widely via websites, web forums, blogs and web-based social networksmission.

- *E.Sowmya is currently pursuing masters degree program in Software neering in SKCET, Coimbatore, India, E-mail: 14mi013@skcet.ac.in*
- R Manikandan, Assistant Professor, *Dept. of computer science engineering, SKCET, Coimbatore, India,E-mail: manikandanr@skcet.ac.in*

2 RELATED WORKS

Katherine Faust et.al [3] discussed the limitations in helpful of a triad census for learning similarities between local structural properties of social networks. The experiential base for this scenario is a collection of 51 social networks computing diverse relational contents such as companionship, recommendation, agonistic encounters, success in fights and dominance relations, amongst a mixture of groups. It discussed about the size, data and density in the social networks. It produced the experimental results for low dimensional dataset and not suitable for high dimensional dataset.

Erwan Le Martelot et.al [4] suggested greedy approach to de-

ISSN 2229-5518

tect the multi scale community group. Community detection is gained more attention in recent years and the proposed work is used greedy method to assess stability as an optimization criterion. It used the Markov time as resolution parameter which is sued to increase the accuracy in the real world network data analysis. The technique is tested against different networks and evaluated to five relevant community discovery algorithms.

David Liben Nowel et.al [5] discussed the link prediction problem for social networks. Link prediction is used to solve the evolution of social network by using important features. This research work analyzes the proximity of modes in the network based on measures. Ben Tasker et.al [6] discussed link prediction in relational dataset. It is focused on the prediction of available links amongst entities in the given domain. The paper proposed the method named as relational markov network to compute the joint probabilistic model on the total link graph. It consists of links, entity and attributes. This approach is increasing the classification accuracy by using relational markov network method.

Darcy Davis et.al [7] proposed a novel probabilistic weighted extension framework for heterogeneous information networks. Link prediction should model the influences among heterogeneous relations and differentiate the configuration mechanisms of every link type, a task which is ahead of the simple topological features frequently used to score potential links in social networks. This scenario is used the supervised methods to improve the system performance but however it produced the loss of information which is caused through the data transformation.

Jan Noessner et.al [8] discussed leveraging terminological structure for object reconciliation. It improves the semantic similarity measure for linked data. The proposed framework permits to merge lexical-a-priori similarities among instances along with the terminological information encoded in this scenario. Jun Huan et.al [9] presented efficient mining of frequent subgraph in the presence of isomorphism. A novel frequent subgraph is introduced in this research to improve the mining efficiency. It is focused on the reduction of number of redundant candidates and increased the substantial performance.

Purnamrita Sarkar et.al [10] proposed two aspects of social network modelling in this research scenario. Initially this approach generalizes a static model of relations to a dynamic model that accounts for friendships drifting over time. Then it demonstrates how to construct it good to find out such models from data when the number of entities n gets large. By using suitable kernel functions the similarity value is increased. Jorge Aranda et.al [11] discussed online social network based recommendation system. This scenario used data from user to user relationships and the problem is addressed in this approach is privacy concept. It shared more useful information to make a clear decision in the large networks. However above mentioned researches have issue with time complexity, inaccurate information, scalability and inefficiency. To overcome all those problems, the proposed work is developed.

Altug Akay et.al [12] proposed network based modelling to improve the healthcare sector using social network information. This research proposed self-organizing map to evaluate correlations among user's posts, positive, negative and neutral opinions. Then this approach builds the model users and their posts. It is used to search more effectively for potential levels in the social network. It discovers dense modules by using a partition stability quality measure. Also it determines the optimal network separation and develops the retrieved modules along with word occurrence information from module-contained user's posts to obtain local and global measures of user's opinion.

3 PROPOSED WORK

The proposed work introduced SOM approach is used to mapping the positive or negative opinion on the drug based on the user posts. Initially we have to preprocess the dataset and perform the stemming as well as stop words.

3.1. Preprocessing and text mining

The main purpose of preprocessing step is to improve the classification performance in the specified dataset. The preprocessing step consist of tokenize, stop words and stemming in this module. This is used to improve the discovering consumer opinion and higher health informatics.

Initially uploaded the data into the first component then the data is processed in the second component ('practice documents to data') by using numerous subcomponents ('mine content', 'tokenize', 'change cases', 'filter stop words', 'filter tokens,' respectively) that removed overload noise (misspelled words and ordinary stop words) to guarantee a consistent set of variables that can be calculated. The last component ('processed information') enclosed the final word list along with every word containing a specific TF-IDF score.We then assigned weights for each of the words found in the user post.

3.2 Cataloging and tagging text data

This approach is used to add a positive tag on negative words and it used the NLTK toolkit for the investigation, and categorization of words to match their accurate meanings within the related settings. For instance, the context must be measured in phrases such as 'I do not feel great' so that the term 'great' will be sufficiently tagged as a negative one (in our case it is tagged as 'great_n' before it is revisited to its detailed location). A sentence that states 'No side effects so I am happy!' resulted in the word 'No' being tagged as 'No_p' (reflecting its positive context) before it is revisited to its particular location. These tagged words are therefore reclassified depends on the context of the post. Then it decreased the number of similar words, both manually (examining the words by using online dictionaries such as Merriam-Webster (http://www.merriam-

webster.com/), and automatically (synonym database software such as the Thesaurus Synonym Database (http://www.language-databases.com/) and Google's synonym search finder.

3.3 Consumer sentiment using SOM

All posts are manually labeled based on the universal user opinion examined in the post as positive, neutral and negative before feeding the composed data for investigative analysis by self-organizing maps (SOM). The manual labeling permitted to use this as a technique of results justification. SOMs are neural networks that create low-dimensional illustration of high-dimensional data. Within this social network, a layer signifies output space along with each neuron allocated a specific weight. The weight values reflect in the cluster and SOM monitors the information to the social network, carrying collectively similar data weights to similar neurons.

While new records are fed into the network, the neighboring weights matching the data change to reflect the new data. This procedure continues until records are no longer fed, resultant in a two-dimensional map. The SOM toolbox utilized and fed with first wordlist TF-IDF vectors. The intention is to evaluate the survival of clusters in the records and how the SOM weights of these clusters will correlate to positive and negative estimations. The SOM is trained by using different map sizes, by using quantization and topographic errors as validation procedures. The previous is the outcome of the average distance among each input and its excellent matching neuron and to calculate how the trained map fits into the input data. The final uses the configuration of the map to protect its topology through representing its accuracy.

3.4. Modeling forum postings using network analysis

In this module, we use network analysis for modeling forum postings. Determining influential users is important process and to achieve the objective, this approach construct networks from forum posts and their replies, when accounting for content-based grouping of posts resulting from the existing forum threads. Networks are composed of nodes and their relations and they are nondirectional (a link among two nodes without a direction) and directional (a link with an origin and an end). The nodal degree of the final measures the number of relations from the origin to the target.

The network-based analysis is extensively utilized in social network analysis depends on its capability to both model and examine intersocial dynamics. We developed a directional network model due to multiple threads along with multiple thread initiators and its internal dynamics between the members reply to thread initiators as well as to other users.

3.5 Identifying subgroups

By using modeling structure, it is easy to identify the subgroups. This modeling frame has accordingly transformed the forum posts into numerous directional networks consist a number of closely associated units. These modules have the feature that they are more solidly linked internally than externally. It decides a multiscale technique that utilizes local and global criterion to recognize the modules, when increasing a separation quality measure named stability. The stability measure assumes the network as a Markov chain along with nodes representing states and edges being probable transitions between these states.

The method is transition probabilities for a random walk of length t (t being the Markov time) allow multiscale analysis. Along with rising scale t and superior modules are found.

The stability of a walk of length t can be expressed as

$$Q_{M_t} = \frac{1}{2m}\sum_{i,j}\left(A_{t_{i,j}} - \frac{d_i d_j}{2m}\right) * \delta(i,j) \qquad (1)$$

where A_t is the adjacency matrix, t is the length of the network, m is the number of edges, i and j are nodes, di is node i's (and j's) strength, and δ (i,j) function becomes one if one of the nodes belong to the same network and zero if it does not belong to any network. At is computed as follows (in order to account for the random walk): $A_t = D \cdot M^t$, where $M = D^{-1} \cdot A$ (D being the diagonal matrix containing the degree vector giving for each node its degree). The method for identifying the optimal modules is based on alternating local and global criteria that expand modules by adding neighbor nodes, reassigning nodes to different modules, and significantly overlapping modules until no further optimization is feasible.

3.6. Module average opinion and user average opinion

The proposed approach is to organize the module average opinions and user average opinions more proficiently. To process the data modules by feeding them with the information gained from the forum posts. In a primary step, it is targeted at recognizing influential users within the networks. Influential users are users which broker most of the information transfer within network modules and whose opinion in terms of positive or negative sentiment towards the treatment is 'spread' to the other users within their containing modules. The TF-IDF scores from the wordlist of positive and negative words and refer table 1 are used to build two forms of measurement. The global measure (pertaining to the whole information module) is represented by the module average opinion (MAO). It examined the TF-IDF scores of postings matching the nodes in a specific module.

MAO $=\frac{sum_+ - sum_-}{sum_{all}}$

$sum_+ = \sum\sum x_{i,j}$ is the total sum of the TF-IDF scores matching the positive words in the wordlist vectors within the module. The units i represent post index. The unit j represents the wordlist index (matching the positive words in the list). $sum_- = \sum\sum x_{i,j}$ is the total sum of the TF-IDF scores matching the negative words in the wordlist vectors within the module. The units i represent post index. The unit j represents the wordlist index (matching the negative words in the list). $sum_{all} = \sum_{i=1}^{N}\sum_{k=1}^{M} x_{ik}$ is the sum of both of the aforementioned

ISSN 2229-5518

sums. The unit k is the index running across variables throughout the entire wordlist.

The local measure that illustrates specific user opinion to each node in the module (the user average opinion, or UAO) that examines the TF-IDF scores to the average of the collected posts of the user is the following:

$$\mathrm{UAO}_i = \frac{\mathrm{Sum}_{i^+} - \mathrm{Sum}_{i^-}}{\mathrm{Sum}_{i\,\mathrm{all}}}.$$

$sum_{i+} = \sum_{j \in P} x_{ij}$ is the TF-IDF score's sum matching to positive words for the ith user's wordlist vector. P is the index set denoting the wordlist's positive variables. $sum_{i-} = \sum_{j \in N} x_{ij}$ is the TF-IDF score's sum matching to negative words for the i-th user's wordlist vector. N is the index set denoting the wordlist's negative variables.

$sum_{all} = \sum_{j=1}^{M} x_{ij}$ is the total of both sums, and j is the index of the whole wordlist.

3.7 EVALUATION RESULT

In this section the existing and the proposed scheme are analyzed by the experimental conclusions. The methods are compared by the metrics such as precision, recall, f-measure and classification accuracy.

Precision

The precision is calculated as follows:

Precision = True positive / True positive+ False positive

Precision is defined as a computation of correctness or quality, whereas recall is a computation of completeness or quantity. And, high precision indicates that the approaches returned significantly more relevant results than irrelevant.

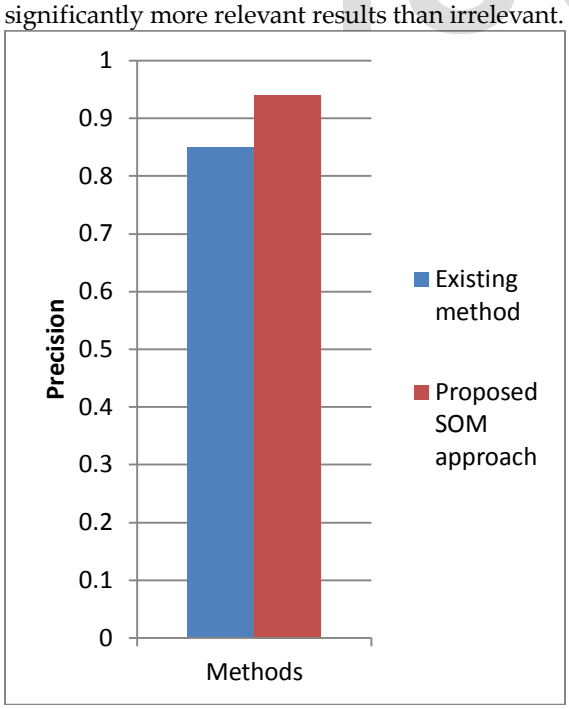

Figure 1. Comparison of Precision

Figure 1 shows the comparison of the existing and the proposed methods based on the precision metric. In x axis the methods are plotted and in y axis the precision ratio is plotted from 0 to 1. The existing method shown lower precision value as 0.85 and the proposed method shown the higher precision values as 0.94. The experimental result concluded that the proposed method provides better precision value than the existing method.

Recall

The calculation of the recall value is done as follows:

Recall = True positive / True positive

Recall is described as the number of relevant documents recovered through a search divided by the total number of accessible relevant documents. Recall is as the number of true positives divided by the total number of elements that essentially belong to the positive class (i.e. the sum of true positives and false negatives, which are items which were not labeled as belonging to the positive class but should have been).

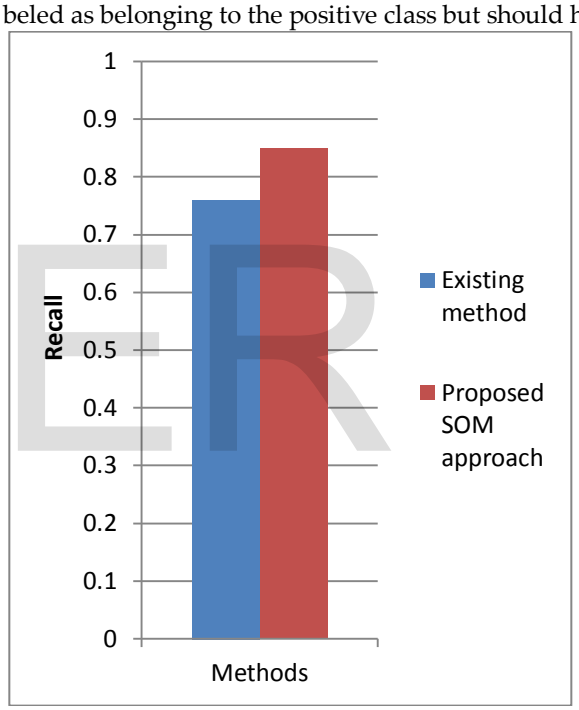

Figure 2. Comparison of Recall

Figure 2 shows the comparison of the existing and the proposed methods based on the recall metric. In x axis the methods are plotted and in y axis the recall ratio is plotted from 0 to 1. The existing method has shown lower recall value as 0.76 and the proposed method has shown the higher recall values as 0.85. The experimental result concluded that the proposed method provides better recall value than the existing method.

F-measure

A measure that combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score:

F = $Precision.Recall \;/\; Precision + Recall$

ISSN 2229-5518

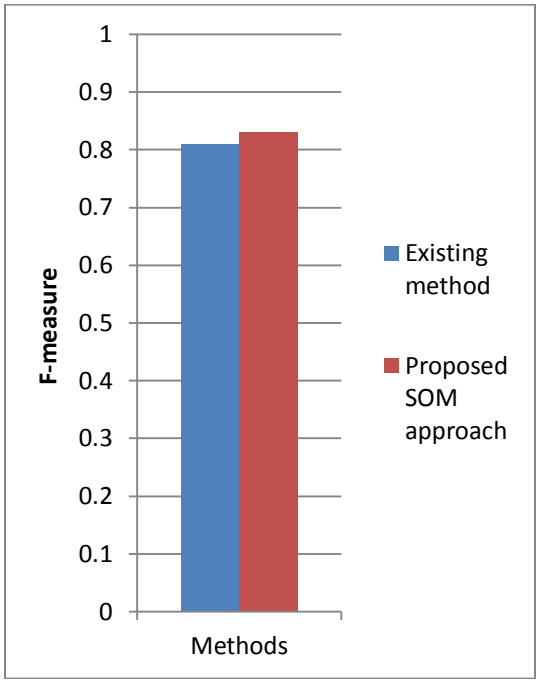

Figure 3. Comparison of F-measure

Figure 3 shows the comparison of the existing and the proposed methods based on the F-measure metric. In x axis the methods are plotted and in y axis the F-measure ratio is plotted from 0 to 1. The existing method has shown lower F-measure value as 0.81 and the proposed method has shown the higher F-measure values as 0.82. The experimental result concluded that the proposed method provides better F-measure value than the existing method.

Accuracy

The accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined.

Accuracy can be calculated from formula given as follows

$$\text{Accuracy} = \text{True positive} + \text{True negative} \;/\; \text{True positive} + \text{True negative} + \text{False positive} + \text{False negative}$$

An accuracy of 100% means that the measured values are exactly the same as the given values.

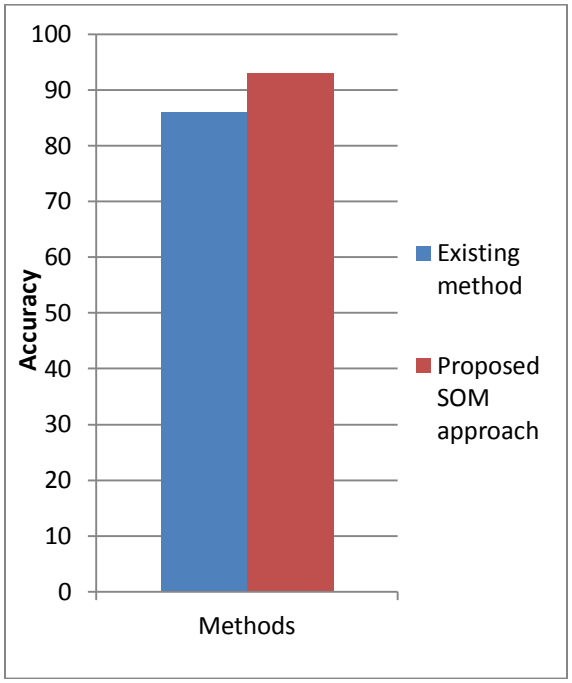

Figure 4. Comparison of Accuracy

Figure 5 shows the comparison of the existing and the proposed methods based on the accuracy metric. In x axis the methods are plotted and in y axis the accuracy value is plotted from 0 to 100. The existing method has shown lower accuracy value as 86 and the proposed method has shown the higher accuracy value as 93. The experimental result concluded that the proposed method provides better accuracy value than the existing method.

4. CONCLUSION

The proposed system introduces a new approach for analyzing the sentiment analysis for lung cancer dataset in social networks. This scenario is able to examine positive and negative sentiment on lung cancer treatment using the drug by mapping the large dimensional data onto a lower dimensional space using the SOM. Many of the user data is grouped into the neighborhood of the map connected to positive opinion, therefore reflecting the universal positive analysis of the users. Consequent network based modeling of the forum capitulated interesting insights on the underlying information exchange among users. The powerfully communication users are recognized by using a multi scale community discovery technique. Furthermore, we are able to recognize possible side effects constantly conversed through groups of users. Such an approach could be used to raise red flags in future clinical surveillance operations as well as highlighting various other treatment related issues. The experimental results concluded that the proposed system is efficient in finding the symptoms and side effects of the drugs based on the data mining approach.

REFERENCES

[1] Griffiths, Frances, et al. "Social networks–The future for health care delivery." Social science & medicine 75.12 (2012): 2233-2241.

[2] Raghupathi, Wullianallur, and Viju Raghupathi, "Big data analytics in healthcare: promise and potential." Health Information Science and Systems2.1 (2014): 3.

[3] Faust, Katherine, "Comparing social networks: size, density, and local structure." Metodoloski zvezki 3.2 (2006): 185.

[4] Le Martelot, Erwan and Chris Hankin, "Multi-scale Community Detection using Stability as Optimization Criterion in a Greedy Algorithm", KDIR, 2011

[5] Liben-Nowell, David, and Jon Kleinberg, "The link-prediction problem for social networks." Journal of the American society for information science and technology 58.7 (2007): 1019-1031.

[6] Taskar, Ben, et al. "Link prediction in relational data" Advances in neural information processing systems. 2003.

[7] Davis, Darcy, Ryan Lichtenwalter, and Nitesh V. Chawla, "Multi-relational link prediction in heterogeneous information networks." Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on, IEEE, 2011.

[8] Ng, Andrew Y., Alice X. Zheng, and Michael I. Jordan, "Stable algorithms for link analysis" Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 2001.

[9] Huan, Jun, Wei Wang, and Jan Prins, "Efficient mining of frequent subgraphs in the presence of isomorphism" Data Mining, 2003, ICDM 2003, Third IEEE International Conference on, IEEE, 2003.

[10] Sarkar, Purnamrita, and Andrew W. Moore, "Dynamic social network analysis using latent space models." ACM SIGKDD Explorations Newsletter7.2 (2005): 31-40.

[11] Aranda, Jorge, et al, "An online social network-based recommendation system." Toronto, Ontario, Canada (2007).

[12] Akay, Altug, Andrei Dragomir, and Bjorn-Erik Erlandsson, "Network-based modeling and intelligent data mining of social media for improving care."Biomedical and Health Informatics, IEEE Journal of 19.1 (2015): 210-218.