www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 5 Issue 12 Dec. 2016, Page No. 19453-19457

Intelligent Data Mining and Network-Based Modeling of Social Media for Improving Care

Mr. Pramod B. Deshmukh[1], *Mrs. Aditi A. Kalia*[2], *Mrs. Vrushali U. Utterwar*[3], *Mrs. Dipali M. Patil*[4]
[1,2,3,4]*Assistant Professor*, D.Y. Patil College of Engineering, Akurdi, Pune, MH, India.
[1]*pramod.deshmukh5@gmail.com* ,
[2]*kalia.86aditi@gmail.com,*
[3]*emailvrushali@gmail.com,*
[4]*patil.dipali41@gmail.com*

Abstract: Intelligently extracting knowledge from social media has newly attracted great interest from the Biomedical and Health Informatics community to simultaneously improve healthcare result and moderate costs using consumer-generated viewpoint. We propose a two-step analysis framework that focuses on positive and negative sentiment, as well as the side effects of treatment, in users' forum posts, and identifies user communities and influential users for the determination of ascertaining user opinion of cancer treatment. We used a Self Organizing Map to analyze word frequency data derived from users' forum posts. We then introduced a novel network-based approach for modeling users' forum interactions and employed a network partitioning method based on optimizing a stability quality measure. This allowed us to determine consumer opinion and analyses influential users within the retrieved modules using information derived from both word-frequency data and network-based properties. Our approach can expand research into intelligently mining social media data for consumer opinion of various treatments to provide speedy, up-to-date information for the pharmaceutical industry, hospitals, and medical staff, on the effectiveness (or ineffectiveness) of future treatments.

Keywords: about four key words separated by commas**.**

1. Introduction

Social media is providing limitless opportunities for patients to discuss their experiences with drugs and devices, and for companies to receive feedback on their products and services [1-3]. Pharmaceutical companies are prioritizing 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 [4]. Social media data harvesting for bio-surveillance have also been reported [5].

Social media enables a virtual networking environment. Modeling social media using available network modeling and computational tools is one way of extracting knowledge and trends from the information 'cloud:' a social network is a structure made of nodes and edges that connect nodes in various relationships. Graphical representation is the most common method to visually represent the information. Network modeling could also be used for studying the simulation of network properties and its internal dynamics.

A sociomatrix can be used to construct representations of a social network structure. Node degree, network density and other large-scale parameters can derive information about the importance of certain entities within the network. Such communities are clusters, or modules. Specific algorithms can perform network-clustering, one of the fundamental tasks in network analysis. Detecting particular user communities requires identifying specific, networked nodes that will allow information extraction. 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.

The nature of social networks makes data collection difficult. Several methods have been employed, such as link mining [6], classification through links [7], predictions based on objects [8], links [9], existence [10], estimation [11], object [12], group [13], and subgroup detection [14], and mining the data [15][16]. Link prediction, viral marketing, online discussion groups (and rankings) allow for the development of solutions based on user feedback.

Traditional social sciences use surveys and involve subjects in the data collection process, resulting in small sample sizes per study. With social media, more content is readily available, particularly when combined with web-crawling and scraping software that would allow real-time monitoring of changes within the network.

Previous studies used technical solutions to extract user sentiment on influenza [17], technology stocks [18], context and sentence structure [19], online shopping [20], multiple classifications [21], government health monitoring [22], specific terms relating to consumer satisfaction [23], polarity of newspaper articles [24], and assessment of user satisfaction from companies [25][26]. Despite the extensive literature, none have identified influential users, and how forum relationships affect network dynamics.

In the first stage of our current work, we employ exploratory analysis using the Self Organizing Maps to assess correlations between user posts and positive or negative opinion on the drug. In a second stage, we model the users and their posts using a network-based approach. We build on our previous study [27] and used an enhanced method for identifying user communities (modules) and influential users therein. The current approach effectively searches for potential levels of organization (scales) within the networks and uncovers dense modules using a partition stability quality measure [28]. The approach enables us to find the optimal network partition. We subsequently enrich the retrieved modules with word frequency information from module-contained users posts to derive local and global measures of users 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 [29].

2. Initial Data Search and Collection

We first searched for the most popular cancer message boards. We initially focused on the number of posts on lung cancer. The chart below gives the number of posts of lung cancer per forum:

Table 1: Initial data

Forums	Posts on Lung Cancer
Cancer-forums.net	36,051
cancerforums.net	34,328
forums.stupidcancer.org	17
csn.cancer.org/forum	7,959

We chose lung cancer because, according to the most recent statistics, it is the most commonly diagnosed cancer in the world for both sexes [30], and the second most prevalent cancer in the US between both the sexes [31][32]. We then compiled a list of drugs used by lung cancer patients to ascertain which drug was the most discussed in the forums. The drug Erlotinib (trade name Tarceva) was the most frequently discussed drug in the message boards. A further search revealed that Cancerforums.net, despite having slightly fewer posts on lung cancer, had more posts dedicated to Erlotinib than the other three message boards mentioned above.

Next, we performed a search of the drug, using both the trade name (Tarceva) and drug name (Erlotinib). The trade name garnered more results (498) compared to the drug name (66). The search using the trade name returned 920 posts, from 2009 to the present date.

3. Initial Text Mining and Preprocessing

A Rapidminer (www.rapidminer.com) [33] data collection and processing tree was developed to look for the most common positive and negative words, and their term-frequency-inverse document frequency (TF-IDF) scores within each post. Figure 1 shows the data collection and processing tree. We initially uploaded the data into the first component (‘Read Excel’). The uploaded data was then processed in the second component (‘Process Documents to Data’) using several sub-components (‘Extract Content’, ’Tokenize’, ’Transform Cases’, ‘Filter Stopwords’, ‘Filter Tokens,’ respectively) that filtered excess noise (misspelled words, common stop words, etc.) to ensure a uniform set of variables that can be measured. The final component (‘Processed Data’) contained the final word list, with each word containing a specific TF-IDF score.

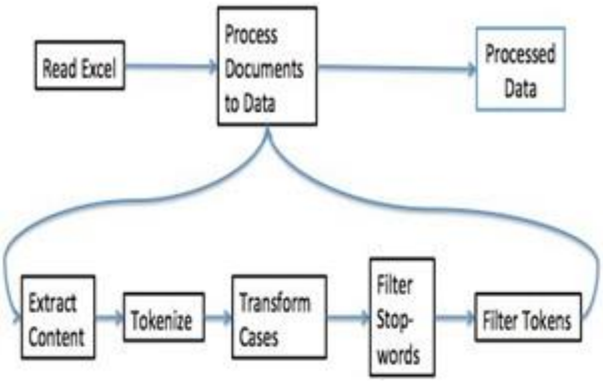

Fig. 1. The processing tree in Rapidminer to ascertain the TF-IDF scores of words in the data

We then assigned weights for each of the words found in the user posts using with the following formula:

$$weight_{t,d} = \begin{cases} \log(tf_{t,d}+1)\log\frac{n}{x_t} & \text{if } tf_{t,d} \ge 1 \\ 0 & \text{otherwise} \end{cases}$$

in which $tf_{i,d}$ represents the word frequency (t) in the document (d), n represents the number of documents within the entire collection, and x_t represents the number of documents where t occurs [30].

4. Information Brokers within the Information Modules

We first ranked individual nodes in terms of their total number of connecting edges (in and out-degree) to identify influential users within the modules.

We then looked nodes in each module based on the following criteria:

1. The nodes have densest degrees within the module (highest number of edges).
2. The UAO scores equate the signs of the MAO of the containing module.

The nodes that qualified were dubbed *information brokers*, based on the above criteria. Their large nodal degrees ensure increased information transfer compared to other nodes while their matching UAO and MAO scores reflect consistency of positive or negative opinion within the containing module.

4.1 Network-based Identification of Side Effects

In the second step of our network-based analysis, we devised a strategy for identifying potential side effects occurring during the treatment and which user posts on the forum highlight. To this goal we overlay the TF-IDF scores of the second wordlist (Table 2) onto modules obtained in Section IIF. The TF-IDF scores within each module will thus directly reflect how frequent a certain side-effect is mentioned in module posts. Subsequently, a statistical test (such as the *t*-test for example) can be used to compare the values of the TF-IDF scores within the module to those of the

DOI: 10.18535/ijecs/v5i12.32

overall forum population and identify variables (side-effects) that have significantly higher scores.

Figure 3 presents a diagram that visually describes the steps in our network-based analysis.

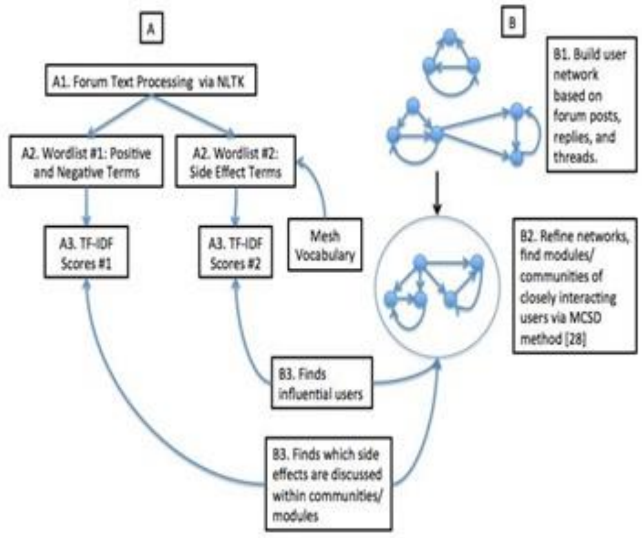
A
A1. Forum Text Processing via NLTK
A2. Wordlist #1: Positive and Negative Terms
A2. Wordlist #2: Side Effect Terms
A3. TF-IDF Scores #1
A3. TF-IDF Scores #2
Mesh Vocabulary
B
B1. Build user network based on forum posts, replies, and threads.
B2. Refine networks, find modules/ communities of closely interacting users via MCSD method [28]
B3. Finds influential users
B3. Finds which side effects are discussed within communities/ modules

Figure 1:. Diagram describing the framework of our network-based analysis. First, the posts collected from the forum via Rapidminer are pre-processed using the NTLK Toolbox (Step A1) and transformed into two wordlists (Step A2). For this step, direct mapping to the MeSH vocabulary is used to identify words representing side-effects Based on the two wordlists, forum posts are transformed into numerical vectors containing word-frequency based TF-IDF scores (step A3). In parallel, forum posts and replies are modeled as a directed network (Step B1). Obtained network is further refined to identify communities/modules of highly interacting users, based on the MCSD method [28] (Step B2). Finally, the two wordlist vectors datasets (their info reflecting the forum information content) are overlaid onto the network modules to identify influential users and highlight side-effects intensively discussed within the modules, respectively (Step B3).

5. Results

Figure 4 shows the unified matrix resulting from the SOM analysis for the wordlist vectors corresponding to the positive and negative terms from the message board Cancerforums.net. A subset consisting of 30% of the data was used for training the SOM. We used a 12 x 12 map size with 110 variables corresponding to the positive and negative terms to ascertain the weight of the words corresponded to the opinion of the drug Erlotinib. As mentioned in the Methods section, each word from the list appeared more than ten times. This achieved a uniform measurement set while eliminating statistically insignificant outliers.

Much of the user's posts converged on three areas of the map. We checked the respective nodes' correlation with their weight vectors' values corresponding to positive or negative words to define the positive and negative areas of the map.

The user opinion of Erlotinib was overall satisfactory, with Table 3 summarizing the satisfaction/dissatisfaction below:

Table 1: User Opinion Of Erlotinib

Satisfaction	Dissatisfaction
70 percent	30 percent
BREAKDOWN OF USER OPINION	
Fully Satisfied (23)	Full Dissatisfaction (4)
Satisfied Despite Side Effects (37)	Dissatisfaction because of Side Effects (20)
Satisfied Despite Costs (10)	Dissatisfaction because of Costs (6)

According to chart, and from our readings of both the user posts and the SOM, the most pressing concern from both camps was the side effects, which are extensively documented in the medical literature.

6. Conclusion

We converted a forum focused on oncology into weighted vectors to measure consumer thoughts on the drug Erlotinib using positive and negative terms alongside another list containing the side effects. Our methods were able to investigate 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. Most of the user data was clustered to the area of the map linked to positive sentiment, thus reflecting the general positive view of the users. Subsequent network based modeling of the forum yielded interesting insights on the underlying information exchange among users. Modules of strongly interacting users were identified using a multi-scale community detection method described in [28]. By overlaying these modules with content-based information in the form of word-frequency scores retrieved from user posts, we were able to identify information brokers which seem to play important roles in the shaping the information content of the forum. Additionally, we were able to identify potential side effects consistently discussed by 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 results have opened new possibilities into developing advanced solutions, as well as revealing challenges in developing such solutions.

The consensus on Erlotinib depends on individual patient experience. Social media, by its nature, will bring different individuals with different experiences and viewpoints. We sifted through the data to find positive and negative sentiment, which was later confirmed by research that emerged regarding Erlotinib's effectiveness and side effects. Future studies will require more up-to-date information for a clearer picture of user feedback on drugs and services.

7. Future Work

Future solutions will require more advanced detection of inter-social dynamics and its effects on the members: such interests of study may include rankings, 'likes' of posts, and friendships. Further emphasis on context posting will require formal language dictionaries that include medical terms for specific diseases, and informal language terms ('slang') to clarify posts. Finally, different platforms will allow up-to-date information on the status of the drug in case one social platform ceases to discuss the drug. Another solution can look at multiple wordlists that can include multiple treatments that, when combined with contextual posting and medical lexical dictionaries, can pinpoint the source (or multiple sources) of user satisfaction (or dissatisfaction), which can open the door towards mapping consumer sentiment of multi-drug therapies for advanced diseases. The combined solutions can open new avenues of post-marketing surveillance research as companies seek real-time, 'intelligent' data of their products and services to remain competitive. This solution can be envisioned on future medical devices that can serve as post-marketing feedback loop that consumers can use to express their satisfaction (or dissatisfaction) directly to the company. The company benefits from real-time feedback that can then be used to assess if there are any problems and rapidly address such problems. Social media can open the door for the health care sector in address cost reduction, product and service optimization, and patient care.

References

[1] A. 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*, Edition of book, Vol. , P. Korosec, , Ed : , p. 267-297, 2010.

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

[3] L. Toldo, "Text Mining Fundamentals for Business Analytics," [25] J. Schectman, (2013, May, 1). Glaxo Mined Online Parent presented at the 11th Annual Text and Social Analytics Summit. Discussion Boards For Vaccine Worries [Online]. Available Boston, MA, 2013.

[4] L. Dunbrack. "Pharma 2.0 – Social Media and Pharmaceutical Sales and Marketing," in Health Industry 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, 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. of the 20th Int. Conf. on Machine Learning (ICML). Washington, D.C., 2003, pp. 496-503

[8] A. Ng, A. Zheng, and M. Jordan, "Stable algorithms for link analysis," in Proc. of the SIGIR Conf. on Information Retrieval. New Orleans, Louisiana, 2001, pp. 258-266.

[9] B. Taskar, M. Wong, P. Abbeel, and D. Koller, "Link Prediction in Relational Data," in Advances in Neural Information Processing Systems (NIPS), Vancouver, B.C., 2003.

[10] D. Liben-Nowell and J.M. Kleinberg, "The link prediction problem for social networks,"Journal of the American Society for Information Science and Technology, Vol. 57, pp. 556-559,May 2007.

[11] Z. Lacroix, H. Murthy, F. Naumann, and L. Raschid, "Links and Paths through Life Sciences data sources," in Proc. of the 1st Int. Workshop on Data Integration in the Life Sciences (DILS), Leipzig, Germany., 2004, pp. 203-211.

[12] 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.

[13] M.E.J. Newman, "Detecting community structure in networks,"European Physical Journal, vol. 38, pp. 321-330, March 2004.

[14] J. Huan and J. Prins, "Efficient Mining of Frequent Subgraphs in the Presence of Isomorphism," in Proc. Of the 3rd IEEE Int. Conf. on Data Mining (ICDM'03), Melbourne, Florida. 2003, pp. 549-552 .

[15] D. Hand, "Principles of Data Mining," Drug Safety, vol. 30, pp.621-622, July 2007.

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

[17] 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, 596-615,Feb. 2010.

[18] S.R. Das and M.Y. Chen, "Yahoo! for Amazon: Sentiment extraction from small talk on the Web," Management Science, vol. 53, pp.1375-1388, Sept. 2007.

[19] E. Riloff, "Little words can make a big difference for text classification," in 18th Annu. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, 1995, Seattle, Washington. pp. 130-136.

[20] W. Yih, P.H. Chang, and W. Kim, "Mining Online Deal Forums for Hot Deals," in WI'04 Proc. of the 2004 IEEE/WIC/ACM Int. Conf. on Web Intelligence, 2004, Beijing, China. pp. 384-390.

[21] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment Classification Using Machine Learning Techniques," in EMNLP'02 Proc. of the ACL-02 Conf. on Empirical Methods in Natural Language Processing, Philadelphia, PA, 2002, pp. 79-86.

[22] X. Feng, A. Cai, K. Dong, W. Chaing, M. Feng, et al., "Assessing Pancreatic Cancer Risk Associated with Dipeptidyl Peptidase 4 Reporting System (FAERS)," J Pharmacovigilance, vol. 1, July 2013.

[23] K.Y. Chan, C.K. Kwong, and T.C. Wong, "Modeling customer atisfactionforproductdevelopmentusinggenetic programming," Journal of Engineering Design, vol. 22, No. 1, pp.56-68, Jan. 2011.

[24] I. Frommholz and M. Lechtenfeld, "Determining the Polarity of Postings for Discussion Search," in LWA 2008-Workshop-Woche: Lernen, Wissen & Adaptivität, Proc., 2008, Würzburg,Germany. pp. 49-56.

[25] J. Schectman, (2013, May, 1). Glaxo Mined Online Parent Discussion Boards For Vaccine Worries [Online]. Available (http://blogs.wsj.com/cio/2013/05/01/glaxo-mined-online-parent-discussion-boards-for-vaccine-worries/)

[26] R. McBride, (2012, August, 1). Merck to Draw on Social Network for patients[Online]. Available (http://www.fiercebiotechit.com/story/merck-draw-social-network-psoriasis-patients/2012-08-13)

[27] A. Akay, A. Dragomir, B.E. Erlandsson,Approach Leveraging Social Media to Monitor Consumer Opinion of Sitagliptin," J. of Biomedical and Health Informatics,in publication .

[28] E. Le Martelot, and C. Hankin, “Multi-Scale Community Detection using Stability as Optimisation Criterion in a Greedy Algorithm,” 2011 Int. Conf. Knowledge Discovery and Information Retrieval (KDIR 2011), Paris, October, pp. 216-225. SciTePress .

[29] National Cancer Institute. (2014, April, 21). Erlotinib Hydrochloride. Available: http://www.cancer.gov/cancertopics/druginfo/erlotini hydrochloride.

[30] National Cancer Institute. (2014, April, 21). Erlotinib Hydrochloride.Available: http://www.cancer.gov/cancertopics/druginfo/erlotini bhydrochloride .

Author Profile

Prof. Pramod B. Deshmukh received the M.Tech degree in Computer Science and Engineering from JNT University Hyderabad and B.E. degrees in Information Technology from Shivaji University Kolhapur in 2014 and 2010, respectively.
He currently is working in D Y Patil College of Engineering, Akurdi, Pune-44.

Prof. Aditi A. Kalia received the M.E. degree in Computer Engineering from Savitribai phule pune University, Pune and B.E. degrees in Computer Engineering from Panjab Technical University, Panjab in 2014 and 2010, respectively.
She currently is working in D Y Patil College of Engineering, Akurdi, Pune-44.

Prof. Vrushali Utterwar received the M.Tech degree in Computer Engineering from Visvesvaraya Technological University, Karnataka and B.E. degrees in Computer Science & Engineering from Sant Gadge Baba Amravati University in 2012 and 2001, respectively.
She currently is working in D Y Patil College of Engineering, Akurdi, Pune-44.

Prof. Dipali M. Patil received the M.E. degree in Computer Engineering from Savitribai phule Pune University, Pune and B.E. degrees in Computer Engineering K. K. Wagh Engineering College, Nashik in 2015 and 2012, respectively.
She currently is working in D Y Patil College of Engineering, Akurdi, Pune-44.