35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

A Novel Data-Mining Platform Leveraging Social Media to Monitor Outcomes of Januvia

A. Akay, *Member, IEEE*, A. Dragomir, PhD., and Björn-Erik Erlandsson, PhD., *Senior Member, IEEE*

***Abstract*—A novel data-mining method was developed to gauge the experiences of the *diabetes mellitus* drug Januvia. Self-organizing maps were used to analyze forum posts numerically to infer user opinion of drug Januvia. Graph theory was used to discover influential users. The result is a word list compilation correlating positive and negative word cluster groups and a web of influential users on Januvia. The implications could open new research avenues into rapid data collection, feedback, and analysis that would enable improved solutions for public health.**

I. 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]. Pharmaceutical companies are prioritizing social network monitoring for their IT departments, potentially creating an opportunity for rapid dissemination and feedback of products and services to optimize and enhance delivery and reduce costs [2].

Existing network model and computational tools (i.e., graph theory) can be used to extract knowledge and trends from the information 'cloud.' Under this paradigm, a social network is a structure composed of interconnected nodes in various relationships such as interests, friendship, kinship, etc. A graphical representation is a common and useful method for visualizing and representing the information.

Network modeling can offer a deeper understanding of social network dynamics. A network model could be used for simulating many network properties such as understanding how users disseminate information among themselves. Another example is studying the enhancement of certain edges of networks (and how certain information affects the enhancements (e.g. how certain user communities evolve based on common interests about specific diseases).

A sociomatrix (or adjacency matrix) is a matrix representing information extracted from social media. It can help construct the network representation. Social networks, though sparse, are leverageable for performing efficient analysis of the constructed networks. Node degree and other large-scale parameters can derive information about the importance of certain entities within the network (drug brands, healthcare providers, pharmaceutical companies, or device manufacturers). Such communities are *clusters*, or *modules*.

Specific algorithms can perform network-clustering, one of the fundamental tasks in network analysis. Finding a community in a social network means identifying nodes that interact with each other more frequently than nodes outside of the group. Community detection can facilitate the extraction of valuable information for the healthcare industry. Pharmaceutical companies could benefit from this for better targeting their marketing spending, and safety issues. Healthcare providers could better understand the level of satisfaction in their services among patients. Doctors could collect important feedback (stored in the labels characterizing these network modules) from other doctors and patients that would help them in their treatment recommendations, and finding adverse events. Lastly, patients could evaluate other consumers' knowledge before making healthcare decisions, resulting in a better-informed and more empowered patient.

Social networks are heterogeneous, multi-relational, and semi-structured, hindering easy data collection. One potential method is link (relationship) mining: it combines social networks, link analysis, hypertext and Web mining, graph mining, relational learning, and inductive logic programming. Researching links involves link-based object classification, object type prediction, group detection, sub-graph detection, and metadata mining. [3]

Traditional social sciences use surveys and involve subjects in the data collection process, resulting in limited raw data (typically hundreds of subjects per study). However, thousands of users of social media produce inordinate amounts of data with rich user interactions. There are two ways to extract this information: 1) crawling using site provided APIs, or 2) scraping needed information from rendered html pages. Many social media sites provide APIs: Twitter, Facebook, YouTube, Flickr, etc. We can also follow how its properties change over time, which would greatly interest public health studies.

II. Methods

We used the self-organizing map (SOM) because of its visual benefits and high-level capabilities that greatly facilitated the high-dimensional data analysis. Bonato et al. has shown how vector quantization algorithms reduce the feature space's size without losing information for identifying clusters in the classification space [4]. We used Graph Theory because of its widespread use in social network analysis, and the ease with which to study, and

A. Akay is with the School of Technology and Health, Royal Institute of Technology, Huddinge, SE-141 52 Sweden (+46 70 190 13 62; e-mail: altu@kth.se).

A. Dragomir, PhD., is with the Department of Biomedical Engineering, University of Houston, Houston, TX 77204-5060 USA (e-mail: adragomir@uh.edu).

B-E Erlandsson, PhD., is with the School of Technology and Health, Royal Institute of Technology, Huddinge, SE-141 52 Sweden (e-mail: bjorn-erik.erlandsson@sth.kth.se).

978-1-4577-0216-7/13/$26.00 ©2013 IEEE

model, user interactions and relationships. We used the directional nodal degree graph because of the nature of the forum and its internal dynamics among the members.

A. Forum Search

We used the forum DiabetesDaily (www.diabetesdaily.com/forum/) based on its popularity. We compiled a list of diabetic drugs and insulin pumps, and separated them by pharmaceutical manufacturer. We chose Januvia based on the large number of posts relating to it compared to the limited number of posts on other drugs and medical devices.

B. Text Mining and Preprocessing

We used Rapidminer (www.rapidminer.com) to convert the search results of the drug Januvia ('positive,' and 'negative') into a network-ranking system reflecting the degree to which the respective network is involved in the opinion formation, and the degree of influence specific users have over the opinion formation, of Januvia.

All Januvia-related posts were put into an Excel spreadsheet and fed into a modified Rapidminer operator decision-tree that removed unwanted characters (HTML tags, punctuation, numbers) and common stop words (e.g. a, the, it, etc.). The process further broke down words into tokens and roots (e.g. working->work; lost, lose->los), resulting in a list with the term-frequency-inverse document frequency (TF-IDF) scores of selected words.

We based this model on how a vector space model can represent text retrieved from the forum: One forum post represents one vector. Each vector component is associated with a particular word. A vector component is assigned a weight, denoting its importance in the document [5]. The weighted components of each vector uses the term-frequency-inverse document frequency (TF-IDF) scheme:

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

where $tf_{t,d}$ is the frequency of word t in document d, n, number of documents in collection, and x_t, number of documents where word t occurs. [6]

We measured the TF-IDF scores of the initial wordlist before splitting it into three word lists (positive words, negative words, and drugs).

C. Wordlists and Self-Organizing Maps

We next fed each word list into a Self Organizing Map (SOM) (www.cis.hut.fi/projects/somtoolbox/) in Matlab to see which vectors clustered together based on the specified words from the initial wordlist. Each feed resulted in different vector groups clustered together. The clusters were checked for vector similarities. Cluster groups containing less than three posts, and no words of interest, were eliminated. The remaining words were counted in the remaining cluster groups.

The SOM is an artificial neural network that produces low-dimensional representation of high-dimensional data. It is a network where a neural layer (projecting the input data) represents the output space, with each neuron corresponding to a cluster with an attached weight vector. The weighted vectors' values reflect the cluster content they are attached to. The SOM presents the available data to the network, linking similar data vectors to the same neurons.

The training process presents new input data to the network that determines the closest weight vector and assigns the data vector to the matching neuron: such neurons (and its neighbors) change to reflect their new value. The neurons farther from the changed neurons rarely change. The process repeats for all input vectors until all convergence criteria are met, resulting in a two-dimensional map. We visually identified subgroups within the map ('positive words' and 'negative words'), and ascertained which posts gravitated towards which words and whether the map reflected consumer satisfaction (or dissatisfaction) towards Januvia.

D. Modeling Forum Postings Using Graph Theory

We modeled information transfer within a graph-based framework to further understand the user interaction dynamics within the forum. A network representing users' interaction was built from users' forum posts (nodes in our network) and their replies (their connecting edges). Seeking to adequately represent the information transfer within the forum, we devised a network model in which we consider posts which are replied (we termed them initial poster nodes, coded with red in Figure 1) most important concerning information content. We added a directed edge from the nodes representing these posts to their replier nodes (Fig 1, coded blue). We further added a reverse node from the replier node to the initial poster node, since direct replies provide a certain amount of information to the initial poster as well (therefore these nodes are linked by the pair of edges with opposite directions). We then add edges to the subsequent posts even if they are informal replies to the initial posters (we termed these as context nodes, coded green). This was done because we consistently observed that subsequent posts still discuss the same thread as the initial post. We set a threshold of 3 to the number context posts considered as influenced by (and thus connected by an edge to) the initial post. Figure 1 shows our information transfer model.

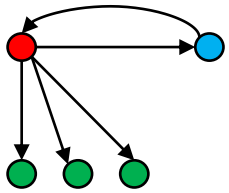

Fig 1. The nodes represent users/posts and the edges represent information among users. Red nodes are initial posters, blue nodes direct replies, green nodes are contextual posts, which receive second order influence.

E. *Identifying sub-graphs*

Our modeling framework has converted the forum posts into a directional graph containing a number of densely connected units (or sub-networks) and unconnected nodes shown in the Figure 2 below:

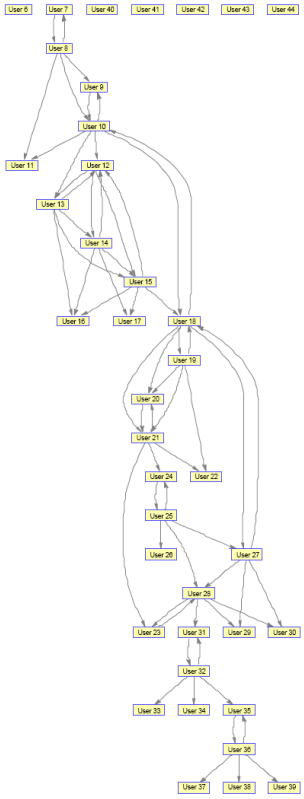

Fig. 2. One of the densely connected sub-networks.

We further pruned the initial sub-networks to identify densely connected components, which we termed *information modules*. Densely connected components of a sub-network contain the maximal group of mutually reachable nodes that do not violate the edge constraints [7]. Fig. 3 shows an example of the information module extracted from the sub-network in Fig. 2. Within each of the modules we computed for each of the nodes, their in- and out-degree, as the total number of incoming and outgoing edges, respectively divided by the total number of edges within the module. Nodes with highest degrees correspond to *influential users*: the node degree is a quantitative measure of the information flow through the node [8].

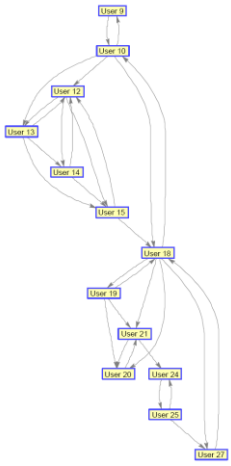

Fig. 3 Information module identified as strongly connected component within the sub-network from Fig. 2.

F. *Module Average Opinion and User Average Opinion*

We further analyzed the information modules by overlapping information collected using the wordlists and TF-IDF scores over them: each module node was attached the vector consisting of the TF-IDF scores corresponding to its respective post. We then defined the Module Average Opinion (MAO) by examining the TF-IDF scores of all posts corresponding to the nodes within the current module as:

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

Where: $Sum_{+} = \sum_i \sum_j x_{ij}$ is the total TF-IDF scores of the wordlist's vector components corresponding to positive words for all users within the current module. The unit i is the node index. The unit j is the wordlist's vector component index (running along all positive words of the wordlist).

$Sum_{-} = \sum_i \sum_k x_{ik}$ is the total TF-IDF scores of the wordlist's vector components corresponding to negative words for all users within the current module. The unit k is the wordlist's vector component index (running along all negative words of the wordlist). $Sum_{all} = \sum_{i=1}^{N} \sum_{l=1}^{M} x_{il}$ is the total of both sums. The unit l represents the index of the whole wordlist.

Similarly, we defined the User Average Opinion (UAO) by examining the TF-IDF scores of the post corresponding to the specific node within the current module:

$$UAO_i = \frac{Sum_{u+} - Sum_{u-}}{Sum_{u\_all}}$$

Where: $Sum_{u+} = \sum_i x_{ui}$ is the total TF-IDF scores of the wordlist's vector components corresponding to positive words for user *u*. The unit *i* is the wordlist's vector component index (running along all positive words of the wordlist). $Sum_{u-} = \sum_j x_{uj}$ is the total TF-IDF scores of the wordlist's vector components corresponding to negative words for user *u*. The unit *j* is the wordlist's vector component index (running along all negative words of the wordlist). $Sum_{u\_all} = \sum_k x_{uk}$ is the total of both sums. The unit *k* represents the index of the whole wordlist.

G. *Information Brokers within the Information Modules*

We characterized both users and their containing information modules using the parameters defined above. We then sought nodes that fulfilled the following criteria:

1. They are influential users within their containing module (nodes with the largest degree ranking as defined in section E)
2. The UAO scores are within the MAO scores (both MAO > 0 and UAO > 0 or both MAO < 0 and UAO < 0).

The second criterion dictates a rule that assumes influential user's opinion is disseminated to the other posters within the densely connected module. We named the nodes fulfilling the above criteria *information brokers*.

III. RESULTS

Figure 4 is the graphical representation of the SOM of the positive words group. We used a 10 x 6 map size with fifty-one variables (positive and negative words chosen from the initial word list) to ascertain how the weight of the words matched the opinion of Januvia. A sizable number of positive words and negative words converged in certain points of the map (the upper and lower portions, respectively). The U-matrix shows that most positive and negative words have converged on opposite sides of the map (with few words on the right and center edge), with slightly heavier negative words. User opinion is roughly divided concerning satisfaction (or lack thereof) of Januvia. The negative opinion stems from either the drug's side effects or discontent with Januvia's performance.

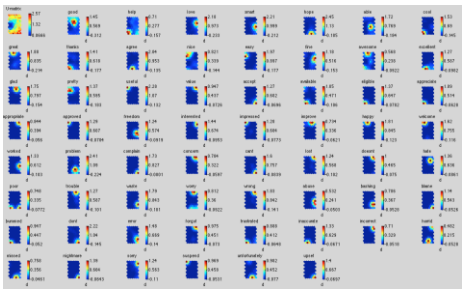

Fig. 4 Positive and Negative Words

We then built the network model using the forum postings that consisted of 711 nodes and 600 edges. Thirty-four information modules containing more than 2 connected nodes were found using our network-modeling framework. We identified influential users by ranking the nodes within each module based on their total number of edges connecting a specific node (in and out-degree) (as described in the methods sections). In Figure 3, the most influential user is user #18 (degree_18=9). The second most influential user is user #12 (degree_12=7).

We next searched the influential users list using the specific criteria define in section G of the Methods. Three users out of the 711 posters were identified using the algorithm as information brokers. The three users' posts reveal that they were mostly active and informative, combining information from sources from the Internet and from personal experience with Januvia. Other members have sought out their wisdom and experience on Januvia. The users' forum posts and interactions have confirmed that these users were the premier information brokers of Januvia in this specific forum.

IV. DISCUSSION

The goal was to transform forum posts dedicated to *diabetes mellitus* into vectors to scan for patterns in the responses to gauge consumer opinion on Januvia. The results open new research avenues into developing a more thorough user-influence web (based on quality of posts and ranking within the forum) algorithms and how that influence (based on quality of posts and ranking within the forum) affects interactions with other users (replies, friendships, post timing and quality). Such advances will be critical in future studies (oncology). The new research avenues will counter this study's limitations (limited number of posts, focus on one drug and on one forum). Social media is becoming an expanding venue for people to express their thoughts, and ideas. It represents a gold mine for companies seeking to optimize health delivery and reduce costs.

REFERENCES

[1] Alberto Ochoa, Arturo Hernández, Laura Cruz, Julio Ponce, Fernando Montes, Liang Li and Lenka Janacek (2010). Artificial Societies and Social Simulation Using Ant Colony, Particle Swarm Optimization and Cultural Algorithms, New Achievements in Evolutionary Computation, Peter Korosec (Ed.), ISBN: 978-953-307-053-7

[2] "Pharma 2.0 – Social Media and Pharmaceutical Sales and Marketing"

[3] Jiawei Han and Micheline Kamber "Data Mining: Concepts and Techniques" 2nd ed., Morgan Kaufmann, 2006

[4] Bonato P, Mork PJ, Sherrill DM, Westgaard RH., "Data mining of motor patterns recordedwith wearable technology," *IEEE Eng Med Biol Mag*.," vol. 22. No. 3, pp. 110-119, May-June 2003] Bonato P, Mork PJ, Sherrill DM, Westgaard RH., "Data mining of motor patterns recordedwith wearable technology," *IEEE Eng Med Biol Mag*.," vol. 22. No. 3, pp. 110-119, May-June 2003

[5] Passmore, D., "*Social Network Analysis: Theory and Applications*"

[6] *Identifying influential users in an online healthcare social network*, X. Tang, C.C Yang, ISI 2010

[7] Aspvall, Bengt; Plass, Michael F.; Tarjan, Robert E. (1979), "A linear-time algorithm for testing the truth of certain quantified Boolean formulas", Information Processing Letters 8 (3): 121–123,

[8] Diestel, Reinhard (2005), Graph Theory (3rd ed.), Berlin, New York: Springer-Verlag