Imperial Journal of Interdisciplinary Research (IJIR)
Vol-2, Issue-5, 2016
ISSN: 2454-1362, http://www.onlinejournal.in

Improving a Novel Data Mining Approach Leveraging Social Media to Monitor Consumer Opinion of Sitagliptin

Sandeep Kaur[1] & Lalitmann Singh[2]
[1]School Of Engineering, Sri Guru Granth Sahib World University, Fatehgarh Sahib.
[2]Assistant Professor, Department Of Computer Science and Engineering, Sri Guru Granth Sahib World University

Abstract: *A novel statistics mining technique changed into evolved to gauge the enjoy of the drug sitagliptin (exchange call Januvia) through patients with diabetes mellitus kind 2. To this purpose, we devised a two-step evaluation framework. Initial exploratory evaluation using self-organizing maps changed into executed to decide structures primarily based on user reviews a number of the forum posts. The consequences have been a compilation of user's clusters and their correlated (fantastic or bad) opinion of the drug. Subsequent modeling the use of community evaluation methods changed into used to decide influential customers among the discussion board members. These findings can open new avenues of studies into speedy facts collection, remarks, and evaluation that can allow advanced consequences and answers for public health and vital feedback for the manufacturer.*

1. Introduction

Online networking, going from individual informing to live foray, is giving boundless chances to patients to talk about their encounters with medications and gadgets. It also provides opportunities for pharmaceutical companies to check the feedback of their drugs and devices. This feedback is very helpful for pharmaceutical companies for example increase in production, increase profit, improve the service and delivery.

Collect the information from social sites and share it in healthcare websites. Therefore, it provides a social networking environment. An appropriate way to extract knowledge and trends from the information 'cloud' would be to model social media using available network modeling and computational tools(such as network-based analysis methods)[1].In this, social network made up of different nodes and edges. Every edge is further connecting to other nodes in various relationships like kinship, friendship and friendship etc.

A grid can speak to data removed from social media (called the socio matrix, or nearness framework) that can construct the system representation. Social networks are very helpful for performing efficient constructed network. In network, parameters are used to drive the information about entities. Such communities called modules or cluster. Clustering is the central important task 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 [1].Community detection can facilitate the extraction of valuable information the whole healthcare industry [1]. It is very beneficial for pharmaceutical companies for their better targeting and improvement.

Social networks are heterogeneous ,multi-relational and semi-structured ,making gathering such data difficult[1].Link mining is one method for combing the different social networks, link analysis, hypertext and web mining, graph mining, relational learning and inductive logic programming[6].Researching links involve several steps:-link-based object classification(categorizes objects based on links and attributes), [7] object type prediction (predicts object types based on attributes ,links, and objects linked to it), [8] link type prediction(predicts the purpose of the link based on the objects involved), [9] link existence prediction (predicts the existence of a link), [10] link cardinality estimation (predicting the number of links (and objects reached) to an object), [11] object reconciliation (determining whether two objects are the same based on their links), [12] group detection (predicting if an object set belongs together),

[13]sub-graph detection (discovering sub-graphs within networks), [14] and metadata mining (mining for data about Data) [15][16].
In this approach three steps:-First, in contrast with already published studies, they identified influential users and to this goal their approach takes into account how forum relationships affects the opinions and behaviors of users [1]. Secondly, they built on the approach[18] and automatically tagged both positive and negative words bases on the context of each posts[1].Lastly ,approach relies on word frequency statistic and employs an exploratory analysis stage, bases on adapted graph theory methods, that aims at identifying user communities and potential influential Users[1].

2. Methods

A. Searching

The first step turned into to find the maximum popular discussion board committed to diabetes mellitus kind 2. We in comparison the quest of four maximum famous diabetes-associated message forums: DiabetesForum.com, Healthboards.com, Foum.lowcarber.org, and DiabetesDaily.com.A list of drugs and devices used by patients changed into compiled and searched in the outcomes of every of the message boards: the purpose become to examine which tablets and/or devices the sufferers were discussing the maximum. Sitagliptin became observed to be the most mentioned drug primarily based on the large variety of posts on the drug on the message boards. The message board DiabetesDaily.com changed into selected due to the fact Sitagliptin changed into the most regularly mentioned drug compared to the alternative four message forums. Within the message board DiabetesDaily.com, a seek using the exchange term 'Januvia' garnered more results (2600 consequences) than the drug time period 'Sitagliptin' (ninety two outcomes). The quested result the usage of 'Januvia' additionally again 713 posts, ranging from 2007 to the present time.

B. Preprocessing

After compiling the posts, they fed the list to modified decision-making tree in Rapid miner (www.Rapidminner.com)[2] to ascertain the most commonly used words[1].The final result of this step is an initial list contain posts with TF-IDF score(term frequency inverse document frequency) and words were then divided into two categories: 'positive' and 'negative.' The weight vector components of each vector (posting) uses the term-frequency-inverse document frequency (TF-IDF)[1].

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

Where $tf_{t,d}$ is the frequency of word *t* in document *d*, *n* number of documents in collection, and t_d number of documents where word *t* occurs [28].

C. Classification

The words with the very best TF-IDF ratings were placed in the discussion board posts and have been then tagged using Python and the NLTK toolkit [29] primarily based at the whether or not they pondered the negativity of a poor phrase and the positivity of a fantastic word based totally on context. as an instance, the term 'I do not feel great' resulted inside the phrase 'incredible' being tagged as 'great__n' before it is lower back to its precise function. Das and Chen used a similar technique in classifying words [30]. We went one step further and added a advantageous tag on bad phrases. A sentence that states 'No facet effects so I am glad!' resulted in the word 'No' being tagged as 'No__p' earlier than it is back to its precise position. These tagged words were then reclassified primarily based on the context of the put up. The following step was to lessen the wide variety of similar phrases. This was executed both manually (checking the phrases the usage of online dictionaries along with Merriam-Webster (http://www.merriam-webster.com/)) and automatically (synonym database software program consisting of word list Synonym Database (http://www.language-databases.com/) and Google's synonym seek finder (using '~' after a word). The wordlist was decreased the usage of the above techniques, ensuing in the improvement of the chart shown under. The chart beneath indicates the phrases divided into the high quality and negative terms. The phrases that companion with advantageous or negative meanings are tagged. This became based on each the frequency of the words used within the forum and the context with which the phrases were used in the posts.
Furthermore, every phrase that seemed less than ten instances becomes additionally eliminated. This allowed us to gain a uniform set of measurements whilst removing statistically insignificant outliers. The stop result changed into a changed wordlist of twenty-eight phrases (fourteen untagged, and tagged advantageous phrases and fourteen untagged, and

tagged poor words) shown in table 1. earlier than feeding the accumulated facts for exploratory analysis via Self Organizing Maps, all posts had been manually labeled according to the overall person opinion discovered inside the submit as wonderful, negative and neutral. The guide labeling swallowed us to use this as a way of results validation.

D. Self-Organizing Map

Self-Organizing Maps (SOMs) are an artificial neural community used for clustering that produces low-dimensional illustration of high-dimensional facts [31]. The SOM is a network in which a neural layer (projecting the enter records) represents the output area, with each neuron corresponding to a cluster with an connected weight vector. The values of the weight vectors reflect the content material of the cluster they're attached to. The SOM provides the available records to the network, linking similar facts vectors to the same neurons. We used the self-organizing map (SOM) due to its visible benefits and excessive-level competencies that significantly facilitated the high-dimensional data analysis. Bonato et al. has proven how vector quantization algorithms reduce the function area's size without losing information for figuring out clusters in the classification space [32]. The training process provides new enter information to the community that determines the closest weight vector and assigns the information vector to the matching neuron: such neurons (and its associates) undergo a variation procedure to mirror their new cost. The neurons further from the modified neurons adapt their weight vectors by using a smaller diploma. The technique repeats for all enter vectors until all convergence standards are met. The end-end result is a two-dimensional map.

We took the modified wordlist and fed it into the SOM toolbox (http://www.cis.hut.fi/tasks/somtoolbox/) in Matlab [33] to see if specific vectors clustered collectively primarily based on the specified phrases from the phrase list. We skilled the SOM with different map sizes, and selected as inner validation measure the quantization and topographic errors. The quantization errors is computed as the average distance between every enter vector and its high-quality matching neuron (BMN) and a degree of the way excellent the educated map suits the Enter data [31]. The topographic mistake considers the map structure and represents the accuracy of the map in Retaining topology. The topographic mistakes value is calculated from the percentage of all facts vectors for which First and second BMNs are not adjoining for measuring topology preservation. The top-quality map length turned into selected based totally at the minimum values of the quantization and topographic mistakes (0.1257 and 10-7 , respectively). The phrase listing vectors have been mapped onto the SOM and emerging clusters have been similarly tested for correlations with effective of poor variables of the word listing vectors. Cluster companies containing three or fewer posts, and no phrases of interest, had been removed. The word occurrences have been counted within the closing cluster groups. We then visually identified subgroups within the map ('tremendous phrases' and 'poor phrases') and ascertained which posts were gravitating closer to which phrases and whether the map pondered patron delight (or dissatisfaction) closer to Sitagliptin.

E. Modeling Forum Postings Using Network Analysis

The subsequent step became to similarly scrutinize the forum posts with the goal of figuring out influential users. To this goal we constructed networks from discussion board posts and their replies. Networks encompass nodes and connections. Networks are either no directional (a connection between factors without a direction) or directional (a connection with a factor of origin to an end). A non-directional nodal diploma measures the variety of connections of a node at the same time as a directional nodal degree measures the range of connections from an original node and its vacation spot(s). Wassermann et al. [34] identified four extraordinary nodes within a community: remoter (connects to no different nodes), Transmitter (connects to different nodes but does no longer obtain them), Receptor (does not hook up with different nodes but receives them), and provider (connects and gets connections). The density of a community measures the present day quantity of (many) connections.

Directional networks divide the maximum number of connections with the number of arrowed connections as shown below:[1]

$$\Delta = \frac{L}{f(f-1)}$$

Where *L* is the number of connections and *f* is the total number of nodes.

For our functions, we used a community-based evaluation technique because of its sizable use in social community analysis, and the benefit with which to look at, and version, consumer interactions and relationships. We used the directional community model because of the character of the forum and its internal dynamics among the members. The technique we chose to construct our community is described in fig 1, which shows how every posting-reply pair is modeled

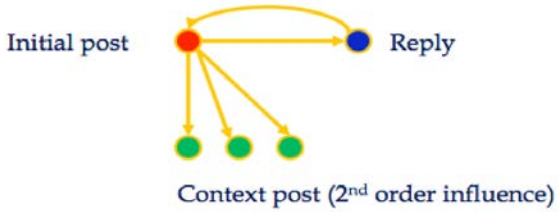

Fig 1. The nodes represent users/posts and the edges represent information among users.[1]

We started out by growing nodes for posts containing direct replies (responses to preceding posts within the discussion board) and delivered bi-directional edges connecting these nodes, as described in Fig 1. The purpose we used bi-directional edges in such instances changed into to reflect the following information transfer (from the preliminary poster to the replier and vice versa, primarily based at the assumption that they both study the initial publish and its reply). Following this, we brought additional edges to the following posts (coded in inexperienced in parent 1). These edges are unidirectional, based totally at the realistic assumption that the subsequent posts persevered to talk about the topic thread (initial publish). We set a threshold of three to the number of subsequent posts which might be taken into consideration as influenced by the preliminary post. This threshold turned into set based totally on our empirical statement of posting contents and their timing.

F. Identifying sub-Graphs

Our modeling framework has consequently transformed the forum posts into several massive directional networks containing some of densely related gadgets (or sub networks) and unconnected nodes proven inside the discern 2 under. We pruned the preliminary networks to discover strongly connected components (or information modules). A strongly related factor is defined for directed networks as a sub-community in which every nodes u and v are linked to every other by way of as a minimum paths (along the connecting edges): one from u to v and one from v to u [35]. The set of rules we used for retrieving those strongly connected components employs a depth-first search approach [36]. Identifying strongly connected additives guarantees that statistics switch inside the sub-community is maximized. Parent three gives the strongly linked element (information module) received from the network in Fig 2.

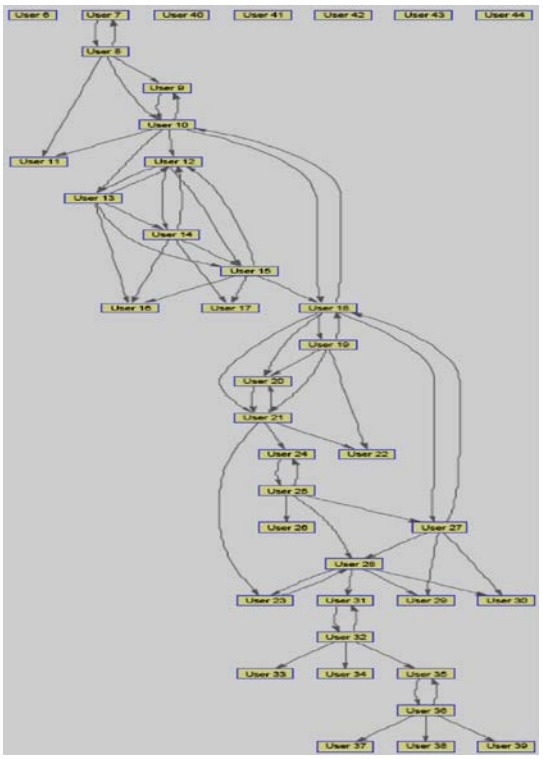

Fig. 2 One of the initial networks build from the Diabetes Daily forum. The forum consisted of 711 nodes, 843 edges, and 34 networks containing more than 2 connected nodes.[1].

G. Module Average Opinion and User Average Opinion

We similarly subtle they obtained information modules by means of enriching them with facts from the posts thru the corresponding phrase listing vectors. At this step, we use the word lists' TF-IDF rankings to derive two measures characterizing user opinion. We first defined a international measure (characterizing the whole records module): the

module common opinion (MAO) via examining the TF-IDF scores of all postings similar to the nodes inside a selected module.
In a comparable manner, we described also a neighborhood degree charactering user opinion (precise to each node in the module), the consumer common opinion (UAO), by way of examining the TF-IDF rankings of the submit similar to the particular node.

H. Information Brokers within the Information Modules

For you to perceive influential customers inside the modules, we first ranked individual nodes in phrases of their total wide variety of connecting edges (in and out-degree). We then looked for nodes within every module that fulfilled the following criteria:
1. They're influential users (nodes with the most important ranges)
2. The UAO ratings are within the MAO ratings (both MAO > 0 and UAO > 0 or both MAO < zero and UAO < zero).
We named the nodes fulfilling the above criteria facts brokers, primarily based on the fact that they possess the best wide variety of connections inside the strongly connected records modules.

2. Literature Review

A. Akay et al.[4] They devised a two-step evaluation framework. Preliminary exploratory evaluation using self-organizing maps became achieved to determine systems based on user evaluations some of the forum posts. The consequences had been a compilation of person's clusters and their correlated (positive or terrible) opinion of the drug. Next modeling the usage of community analysis strategies became used to decide influential users a number of the forum participants. Those findings can open new avenues of research into fast statistics collection, remarks, and analysis which could permit stepped forward outcomes and answers for public fitness and crucial comments for the producer
Zoe Lacroix et al.[5]. They model the records items in those resources and the hyperlinks between objects as an object graph. They become aware of a set of thrilling houses for hyperlinks and paths, including out degree, photo of a hyperlink, cardinality of statistics gadgets and hyperlinks, the variety of wonderful items reached through some links, etc. Analogous to database cost fashions, we use information from the item graph to expand a framework to estimate the end result length for a query at the item graph. Analogous to schooling and testing, we use sampled facts from queries to estimate the end result length. They validate our models using data sampled from 4 NIH/NCBI information sources. Their study presents a foundation for querying and exploring statistics assets.

Noessner, Jan, et al [7]. They endorse a singular technique to item reconciliation that is primarily based on an existing semantic similarity degree for connected statistics. They adapt the degree to the item reconciliation hassle, present precise and approximate algorithms that successfully implement the techniques, and provide a scientific experimental evaluation based on a benchmark dataset. As their predominant end result, they show that using lightweight ontology's and schema facts substantially improves object reconciliation within the context of connected open statistics.

Noessner, Jan, et al [7]. They increase the analysis and display the way it gives insight into approaches of designing solid hyperlink analysis techniques. This in flip motivates new algorithms, whose overall performance they examine empirically the usage of citation facts and web hyperlink statistics.

Lise Getoor, Christopher P. Diehl [68]. Numerous datasets of interest today are best portrayed as a connected gathering of interrelated items. These might speak to homogeneous systems, in which there is a solitary article sort and connection sort, or wealthier, heterogeneous systems, in which there might be different question and connection sorts (and conceivably other semantic data). Illustrations of homogeneous systems incorporate single mode informal organizations, such as individuals associated by kinship joins, or the WWW, a gathering of connected site pages. Illustrations of heterogeneous systems incorporate those in restorative areas portraying patients, ailments, medicines and contacts, or in bibliographic areas portraying productions, creators, and venues. Join mining alludes to information mining procedures that expressly consider these connections when building prescient or clear models of the connected information. Generally tended to connection mining errands incorporate article positioning, bunch recognition, aggregate classication, join forecast and sub graph revelation. While system investigation has been concentrated on top to bottom specifically territories for example, interpersonal organization examination, hypertext mining, and web investigation, just as of late has there been a cross-preparation of thoughts among these different groups. This is an energizing,

quickly extending region. In this article, we survey a portion of the normal rising subjects.

3. Conclusion

Previous to the final SOM, a subset of the data becomes used for training the SOM. This turned into to ensure that SOM became trained to as it should be model a pattern set of the records prior to receiving the whole information set. To this quit, thirty percent of the records have been selected for education the SOM. We used a 13 x 13 map size with twenty-eight variables from the changed wordlist to examine the weight of the words corresponded to the opinion of the drug Sitagliptin. A criterion for choosing the variables become that every word have to appear ten times and above. This allowed us to attain a uniform set of measurements whilst eliminating statistically insignificant outliers. The bulk of the person's posts converged on four points of the map. We checked the correlation of the respective nodes with the values in their weight vectors corresponding to positive or terrible words. This is how we defined the high quality and negative regions of the map.

An image starts off evolved to emerge of user opinion that is more or less divided on the subject of satisfaction (or lack thereof) of the drug Sitagliptin. One source of bad opinion stems from the aspect results of the drug. A evaluate of the clinical literature has confirmed the very equal side effects that the customers were discussing [37-40]. Different resources of bad opinion range from consumer frustration of the drug costs to frustration on the clinical network. Superb opinions in particular stemmed from pleasure by using users who switched to it based on suggestions from a physician. The SOM evaluation meditated on the tough department of person opinion of Sitagliptin at the forum, based at the reasons said above.

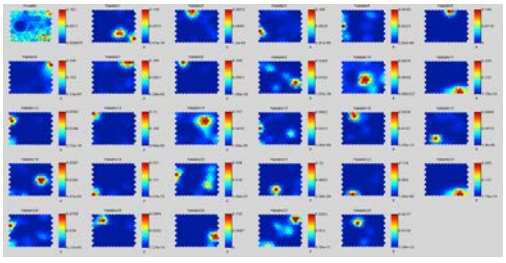

Fig3. Results of the SOM analysis on posts from the Diabetes Daily forum. Top left panel shows the unified distance matrix in which several user clusters can be observed. The rest of the panels display individual word list[1].

The following step changed into to become aware of precise, influential users within the forum. On the Diabetes Daily forum, six users out of the 711 posters were diagnosed as data brokers as proven in parent 5 under. The figure suggests the facts modules wherein those customers are living. The densities of these modules range from 0.25 (for module containing person 18 Pinnacle right panel in Fig 4)

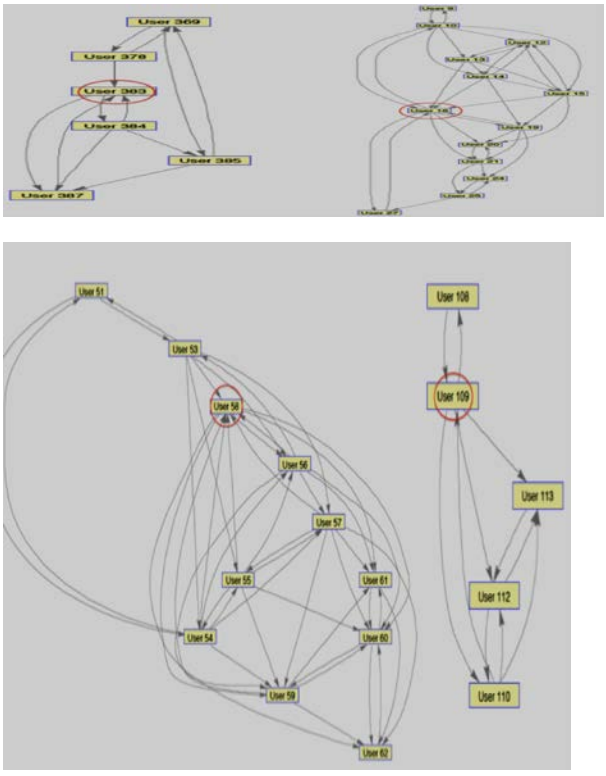

Fig. 4 *Six users were identified as information brokers on the Diabetes Daily Forum. Modules in which these six users reside are shown in this figure. Network densities for the displayed modules are: 0.40 (top left module), 0.25 (top right), 0.42 (mid left), 0.40 (mid right), 0.42 (bottom left) and 0.55 (bottom right)[1].*

To 0.55 (for modules containing person #109, backside proper panel in determine five). These density values are in the observed density values c programming language (closer to the higher limit), while in comparison to those generally mentioned in social networks, thus confirming our network modeling method [41-42]. Those density values are enormously high whilst as compared to those generally observed In social networks [41]. The

information brokers (turned around in crimson) we retrieved had been categorized as carriers primarily based on Wassermann et al.'s strategies. They received, and connected to, different nodes inside the community and their connections have been the densest. The directional nature of the networks represents the extent of interaction among the companies and different customers. A radical studying of the posts of these six users revealed that they have been usually informative, combining facts from assets from the internet and from non-public enjoy with Sitagliptin. Their information and enjoy concerning Sitagliptin turned into positively received and well-liked by way of other individuals. Those users had been also lively in answering questions that other customers (from novices to lengthy-time contributors) had concerning Sitagliptin. Their discussion board 'behavior' has confirmed to us that these customers were the most fulfilling information brokers of the medicine Sitagliptin on the Diabetes Daily discussion board.

4. References

[1] A. Akay, Member, IEEE, A. Dragomir, Member, IEEE, Björn-Erik Erlandsson, Senior Member, IEEE.

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

[3] L. Toldo, "Text Mining Fundamentals for Business Analytics, "Presented at the 11th Annual Text and Social Analytics Summit. Boston, MA, 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*.M 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 Sub graphs 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 Inhibitors: Data Mining of FDA Adverse Event Reporting System (FAERS)," *J Pharmacovigilance*, vol. 1, July 2013.

[23] K.Y. Chan, C.K. Kwong, and T.C. Wong, "Modeling customer satisfaction for product development using genetic 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-onlineparent-discussion-boards-for-vaccine-worries/)

[26] R. McBride, (2012, August, 1). *Merck to Draw on Social Network for Psoriasis Patients* [Online]. Available (http://www.fiercebiotechit.com/story/merck-draw-socialnetwork-psoriasis-patients/2012-08-13)
[27] I. Mierswa, M. Wurst, W. Michael, R. Klinkenberg, M. Scholz, and T. Euler, "YALE: Rapid Prototyping for Complex Data Mining Tasks," in *Proc. of the 12th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD-06)*, 2006, Philadelphia, PA. pp. 935-940
[28] P. Soucy and G.W. Mineau, "Beyond TFIDF Weighting for Text Categorization in the Vector Space Model," *IJCAI'05 Proc. Of the 19th Int. Joint Conf. on Artificial intelligence*, 2005, Edinburgh, Scotland, UK. pp. 1130-1135
[29] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. Sebastopol, CA: O'Reilly Media, 2009, pp. 504
[30] 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.
[31] T. Kohonen. *Self-Organizing Maps*, 3rd ed. Heidelberg-Berlin:
Springer, Dec. 2000.
[32] P. Bonato, P.J. Mork, D.M. Sherill, and R.H. Westgaard, "Data mining of motor patterns recorded with wearable technology". *Eng. in Med. and Biol. Mag., IEEE*, vol. 22, pp. 110-119, May- June 2003.
[33] J. Vesanto, J. Himberg, E. Alhoniemi, J. Parhankangas, "Self- Organizing Map in MATLAB: The SOM Toolbox," *Proc. of the Matlab DSP Conf.*, 1999, Espoo, Finland. pp. 35-40
[34] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*. New York: Cambridge University Press, 1994, pp. 825
[35] Ibid
[36] T.H. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein, *Introduction to Algorithms*, 2nd. Cambridge, MA: MIT Press and McGraw-Hill, 2001
[37] A.V. Matveyenko, S. Dry, H.I. Cox, et al., "Beneficial Endocrine but Adverse Exocrine Effects of Sitagliptin in the Human Islet Amyloid Polypeptide Transgenic Rat Model of Type 2 Diabetes Interactions with Metformin," *Diabetes*, vol.
58, pp. 1604-1615, July 2009.

[38] S. Singh, H. Chang, T.M. Richards, et al., "Glucagon like Peptide 1–Based Therapies and Risk of Hospitalization for Acute Pancreatitis in Type 2 Diabetes Mellitus: A Population-Based Matched Case-Control Study," *JAMA Intern Med.*,vol. 173, pp. 534-539, Feb. 2013
[39] M. Elashoff, A.V. Matveyenko, B. Glier, et al., "Pancreatitis, Pancreatic, and Thyroid Cancer with Glucagon-Like Peptide-1– Based Therapies." *Gastroenterology*, vol. 141, pp.150-156, July 2011
[40] S. Shimoda, S. Iwashita, S. Ichimori, et al., "Efficacy and safety of sitagliptin as add-on therapy on glycemic control and blood glucose fluctuation in Japanese type 2 diabetes subjects ongoing with multiple daily insulin injections therapy." *Endocrine Journal*, vol. 60, No. 10, pp.1207-1214, Aug 2013
[41] K. Faust. "Very local structure in social networks," *Sociological Methodology*, vol. 37, pp. 209-256, Nov. 2007
[42] K. Faust. "Comparing social networks: Size, Density and Local Structure," *Advances in Methodology and Statistics*, vol. 3, No.2, pp.185-216, 2006.
[43] D. Jensen and H. Goldberg. AAAI Fall Symposium on AI and Link Analysis. AAAI Press, 1998.
[44] Jason Ong, Syed Sibte Raza Abidi. In *International Conference on Artificial Intelligence (IC-AI'99), June 28-July 1 1999, Las Vegas*