ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 5, Issue 4, April 2016

Improving Health Care in Social Media Using Data Mining

I.Sahaya Jasmine Jenifer, Dr.K.Kavitha

Abstract— **Social media becomes an increasingly essential tool for reaching patients and it allows to interact with patient population in real time on a more private level. Patients engage with health care organizations on social media with the anticipation that they will receive a friendly and helpful response. But to respond quickly and accurately, it's best to have responses to address common questions, comments and concerns. This paper describes some data mining techniques which concerns text mining to analyze unstructured text content. Text and structural data mining of network and social media provides a new disease surveillance resource and can identify online communities for targeted open health communications to promise wide dissemination of pertinent information. Text mining is shown to recognize trends in flu posts that associate to real-world influenza -like illness patient report data. Social media, a communication boon for the public health community has the potential to support and change many health-related behaviours and issues particularly in times of crisis.**

Index Terms— Data mining, health informatics, graph-based data mining, web and social media, social network analysis.

I. INTRODUCTION

Data mining, or knowledge detection, is the computer-assisted procedure of quarrying through and analysing enormous sets of data and then extorting the meaning of the data. Data mining tools predict behaviours and future trends, allowing businesses to construct practical, knowledge-driven decisions. Data mining tools can answer business questions that were traditionally too time consuming to determine [10]. They search databases for hidden patterns, finding predictive information that specialists might miss because it lies outside their expectations. Data mining obtain its name from the similarities among searching for valuable information in a large database and mining a mountain for a vein of valuable one. Both processes need either sifting through an huge amount of material, or intelligently probing it to find where the value resides. Though data mining is unmoving in its immaturity, companies in a wide range of businesses – including economics, health care, manufacturing, transportation,– are already using data mining tools and schemes to obtain advantage of past data. By using pattern appreciation technologies and statistical and mathematical methods of filter through warehoused information, data mining helps forecasters recognize significant facts, relationships, trends, exceptions, and anomalies that may or else go unobserved.

It allows these companies to decide relationships between "internal" elements such as price, product positioning, or workers skills, and "external" elements such as economic pointers, competition, and consumer demographics. And, it allows them to determine the contact on sales, customer satisfaction, and corporate profits[10].

Finally, it permits to "drill down" into précis information to view detail transactional data. With data mining, a merchant could employ point-of-sale records of customer purchases to send targeted promotions based on an individual's purchase history. Data mining damaged in many areas, in particular health care communities. Detecting particular user communities involves identifying specific, networked nodes that will allow information extraction [10].

II. EXISTING SYSTEMS

Social media communication is an ever more utilized outlet for people to freely create and post information that is dispersed and consumed worldwide through the Internet. News media, traditional scientific outlets, and social media make a platform for minority viewpoints and personal information, which is not being captured by other resources.

Social media can make a sense of anonymity, allowing for unadulterated personal appearance when compared to traditional face-to-face meetings, particularly among young people and about intimate matters [11]. In this esteem, social media provide an additional familiar source of data that can be used to identify health information not accounted to medical officials or health groups and to reveal viewpoints on health-related topics, especially of a sensitive scenery.

I.Sahaya Jasmine Jenifer *, M.Phil Research Scholar, Department of Computer Science, Mother Teresa Women's University, Kodikanal , India,*
Dr.K.Kavitha*, Assistant Professor Department of Computer Science, Mother Teresa Women's University, Kodikanal , India.*

In the past 10 years, investigate articles connecting disease surveillance with Internet use have increased in number, mainly likely due to the enlarge in availability of health-related information from various Internet sites. For instance, Wikipedia article hits [2], Google search provisions (Google Flu Trends) [3], [4] were modeled alongside the number of patients with influenza-like illness (ILI) accounted by the Centers for Disease Control and Prevention (CDC). Several literature analysis have looked at the potential of this sort of research to benefit human health.

Moorhead et al. conducted a appraisal of research studies to recognize potential uses, benefits, and limitations of social media to connect the general public, patients, and health experts in health communication [5]. Although objects identified advantage from using social media in health communications, the writers note a lack of research focused on the valuation of short- and long-term impacts on health announcement practices. Bernardo et al. provided a scoping assessment of the use of search queries and social medium in disease surveillance [6]. First reported in 2006, the reviewed journalism highlighted accuracy, speed, and cost performance that was as good as to existing disease surveillance systems and recommended the exercise of social medium programs to support those systems.

Velasco et al. defined their literature examination to include only peer-reviewed articles on event-based disease observation [7] in which they recognized the existing systems. Walters et al. described many systems implemented and dedicated to bio surveillance, defined as "the discipline in which diverse data streams such as these are described in real or near-real time to offer early warning and situational awareness of events disturbing human, plant, and animal health," many of which within roughly human disease outbreaks [8].

The paper points out that including promising medium, such like blogs and Short Message Service (SMS), into these schemes along with standardized metrics to calculate the performance of different surveillance systems is critical to the advancement of these early warning systems.

As elements of the International Society for Disease Surveillance (ISDS), recognized a social media effective group (henceforth called the workgroup) to expand research, technology, and operational innovations in electronic public health surveillance.

Now proposed to evaluate the use of social media to allow public fitness professionals to realize positive, valuable, and apt community health outcomes at the local, point, regional, national, and global levels. To address these goals, followed the PRISMA(Preferred Reporting Items for Systematic Reviews and Meta-Analyses)procedure [8] by systematically compiling and analyzing copy that demonstrates improvement in electronic public health observation through the use of social media.

By focusing on how investigation on social media data can be used for actionable disease observation, are able to bring to light the most excellent ways of using these tools to target vulnerable populations and get better public health in the broad spectrum from identifying and checking disease occurrences to addressing traditionally difficult health concerns, such as young adults drug and alcohol use.

III. DESIGN OBJECTIVES OF PROPOSED SYSTEM

The main design objective of this work includes:

- Proposed Design is able to convert a forum into weighted vectors to calculate consumer thoughts using positive and negative conditions alongside another list containing the side effects
- The proposed methods is able to find positive and negative sentiment by mapping the huge dimensional data onto a lower dimensional space using the SOM.
- Provides more security by distributing credentials to the users.

IV. SYSTEM MODEL

The System model consists of the User/Patient, Physician, and Admin. It contains the following,

1. Provides security by using identification
2. All the three entities(Patient,Physician,Admin) can view other's post
3. Provides interaction using reply button
4. Reduce search time by providing instant search
5. Unlimited Accessibility

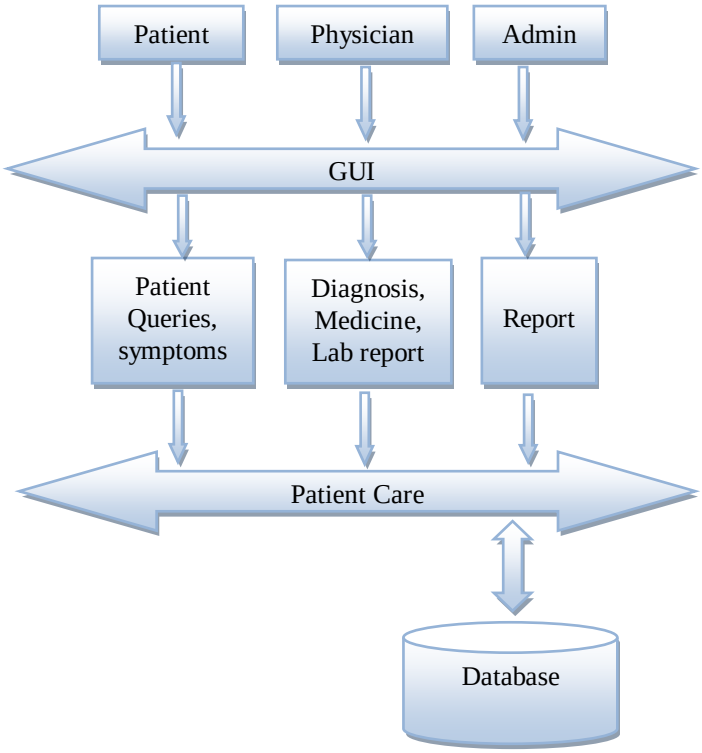

Fig1. System Architecture

According to the Figure, a forum consist of patient, doctor, admin. All these three entities will enter into the forum by using their credentials. Once they entered the patient can post their queries, Admin will follow all the users/patients. The physician role is to give/suggest treatment for their illness. All the details are stored in a database for later use.

V. METHODOLOGY

A. Preliminary Data Search and Compilation

First the User/Patient or Physician searched for the most popular message boards.

B. Text Mining and Preprocessing

Data collection and development tree was developed to look for the mainly common positive and negative words, their term-frequency-inverse document frequency (TF-IDF) scores in each post.[1]

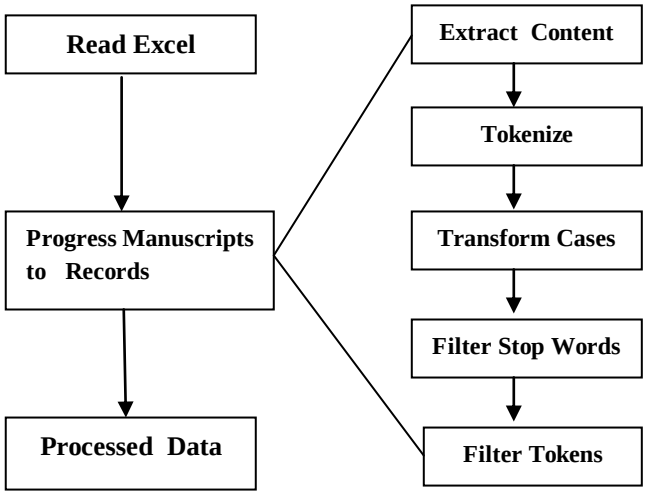

Fig. 2 The data collection and processing tree

Firstly, upload the data into the first part ('Read Excel'). The uploaded data is then processed in the second part ('Progress Manuscripts To Records) using several subparts ('Extract Content', 'Tokenize', 'Transform Cases', 'Filter Stopwords', 'Filter Tokens,' respectively) that filtered surplus noise (misspelled words, common stop words, etc.) to make sure a uniform set of variables that can be measured. The final part ('Processed Data') comprised the ultimate word list, with each word containing a specific TF-IDF score[1].

Then allocated weights for each of the words found in the user posts using by the following formula[1],[12]:

$$weight_{t,d} \begin{cases} log\ t\ f_{i,d} + 1)\ log\ n/x_i & \text{if } t\ f\ t,d \geq 1 \\ 0 & otherwise \end{cases}$$

in which *tfi,d denotes* the word frequency (*t*) in the document (*d*), *n* denotes s the number of papers within the entire collection, and *xt* represents the number of papers where *t* occurs [1],[14].

C. User/Patient Opinion Using a SOM

For this part of the analysis, SOM(Self Organizing Map) make low-dimensional representation of high-dimensional data [12]. The purpose of using this to evaluate the existence of clusters in the data and how the SOM weights of these clusters would compare to positive and negative opinion[1]

D. Modeling Forum Postings Using Network Analysis

The network-based analysis is broadly used in social network This is based on its ability to both model and analyze inter social dynamics[1]. To this goal, built a network from forum posts and their replies, as accounting for content-based grouping of posts outcome from the existing forum outfits .First, the posts collected transformed into two wordlists. Based on the two wordlists, forum posts are modified into numerical vectors containing word-frequency based TF-IDF scores In parallel, forum posts and replies are modeled as a directed network. Obtained network is further refined to recognize communities/modules of highly interacting users, based on the some methods[15].

F. Categorizing Sub graphs

Modeling framework has consequently converted the forum posts into several huge directional networks containing a number of compactly connected units (or modules).These modules have the characteristic that they are more compactly attached internally than externally. Chose a multi-scale method that uses local and global criterion for identifying the modules, while maximizing a divider quality measure called stability[1],[13].

The stability measure believes the network as a Markov chain, through nodes representing states and edges being possible conversions among these conditions. In the writers proposed an approach in which changeover probabilities for a arbitrary walk of length *t* (*t* being the Markov time) allow multiscale analysis. With increasing scale *t*, larger and larger modules are found[13].

E. Network-Based Classification of Side Effects

In the second step of network-based analysis, developed a strategy for identifying potential side effects occurring in the treatment and which user posts on the forum show up. To this goal, overlay the TF-IDF scores of the second wordlist onto modules are obtained. The TFIDF scores within each module will thus directly reflect how common a certain side-effect is cited in module posts. Then, a statistical test can be used to evaluate the values of the TF-IDF scores within the component to those of the overall forum population and identify variables (side-effects) that have considerably higher scores[1].

VI . CONCLUSION

The workflow of analyzing healthcare content in the social media helps to overcome the limitations of large scale data analysis and manual analysis of user generated textual content in social media. This work can help the users to be updated with the effectiveness of the medicines and it can even suggest them with few better medications available. This work also provides feedback to the healthcare system organization and pharmaceutical companies for the available treatments and medicines.

REFERENCES

[1] Altug Akay, Andrei Dragomir, Bj¨orn-Erik Erlandsson. Network-Based Modeling and Intelligent Data Mining of Social Media for Improving Care. IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, VOL. 19, NO. 1, JANUARY 2015

[2] McIver D, Brownstein J. Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the United States in Near Real-Time. PLoS Comput Biol. 2014;10: 1–8.doi: 10.1371/ journal.pcbi. 1003581 .

[3] Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009;457: 1012–1014. doi: 10.1038/nature07634. pmid:19020500.

[4] Nsoesie E, Buckeridge D, Brownstein J. Guess Who's Not Coming to Dinner? Evaluating Online Restaurant Reservations for Disease Surveillance. J Med Internet Res. 2014;16. doi: 10.2196/jmir.2998

[5] Moorhead S, Hazlett D, Harrison L, Carroll J, Irwin A, Hoving C. A New Dimension of Health Care: Systematic Review of the Uses, Benefits, and Limitations of Social Media for Health Communication. J Med Internet Res. 2013;15. doi: 10.2196/jmir.1933.

[6] Bernardo TM, Rajic A, Young I, Robiadek K, Pham MT, Funk JA. Scoping Review on Search Queries and Social Media for Disease Surveillance: A Chronology of Innovation. J Med Internet Res. 2013;15: e147. doi: 10.2196/jmir.2740. pmid:23896182

[7] Acera E, Agheneza T, Denecke K, Kirchner G, Eckmanns T. Social Media and Internet-Based Data in Global Systems for Public Health Surveillance: A Systematic Review. Milbank Q. 2014;92: 7–33. doi: 10.1111/1468-0009.12038. pmid:24597553

[8] Walters R, Harlan P, Nelson N, Hartley D. Data Sources for Biosurveillance. Wiley Handbook of Science and Technology for Homeland Security. John Wiley & Sons, Inc.; 2009. pp. 1–17.

[9] Moher D, Liberati A, Tetzlaff J, Altman D. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med. 2009;6. doi: 10.1371/journal.pmed.1000097

[10]http://www.csbdu.in/csbduold/pdf/Introduction%20to%20Data%20Mining%20and%20its%20Applications.pdf

[11] Iafusco D, Ingenito N, Prisco F. The chatline as a communication and educational tool in adolescents with insulin-dependent diabetes: preliminary observations. Diabetes Care. 2000;23: 1853–1853. doi: 10.2337/diacare.23.12.1853b.

[12] I. Mierswa, M. Wurst, W. Michael, R. Klinkenberg, M. Scholz, and T. Euler, "YALE: Rapid prototyping for complex data mining tasks," in Proc. 12th ACM SIGKDD Int. Conf. Knowl. Disc

[13] E. Le Martelot and C. Hankin, "Multi-scale community detection using stability as optimisation criterion in a greedy algorithm," Proceedings of the 2011 International. Conf. erence on Knowledge Discovery and Information Retrieval (KDIR 2011), Paris, France: SciTePress, Oct. 2011, pp. 216–225.

[14] World Cancer Research Fund International. (2013, Dec. 13). Cancer Statistics Worldwide. [Online]. Available: http://www.wcrf.org/ cancer_statistics/world_cancer_statistics.php

[15] S. Wasserman and K. Faust, Social Network Analysis: Methods and Applications. New York, NY, USA: Cambridge University Press, 1994, pp.825