A Survey and Analysis of Various Health-Related Knowledge Mining Techniques in Social Media

D. Krithika Renuka
Asst Professor
Dept of Computer Science (PG)
PSGR Krishnammal College for Women
Coimbatore.

B. Rosiline Jeetha, PhD
HOD, Professor
Dept of Computer Science
Dr.N.G.P College of Arts and Science
Coimbatore

ABSTRACT
Smart extraction of knowledge from social media has received the recent interest of the Biomedical and Health Informatics community for the simultaneous improvement of healthcare outcomes and lessen the expenses making use of consumer-generated reviews. Social media provides chances for patients and doctors to share their views and experiences without any obtrusion through online communities that might generate information, which is much beyond what is known by the domain experts. Nonetheless, for conventional public health surveillance systems, it is difficult to detect and then monitor the concerns related to health and the changes seen in attitudes of the public towards health-related problems. To solve this problem, several studies have shown the usage of information in social media for the discovery of biomedical and health-related information. Several disease-specific knowledge exchanges are now available on Face book and other portals of online social networking. These kind of new sources of information, support, and engagement have gone to become significant for patients who are suffering with the disease, and still the quality and the content of the knowledge contributed in these digital areas are not properly comprehended. The existing research methodologies are discussed with their merits and demerits, so that the further research works can be concentrated more. The experimental tests conducted were on all the research works in mat lab simulation environment and it is compared against each other to find the better approach under various performance measures such as Accuracy, Precision and Recall.

Keywords
Social media, Health related issues, Sentiment Classifications and SOM.

1. INTRODUCTION
In recent years, the rapid emergence of social media services such as Face book and Twitter allows more and more users to participate in online social activities such as posting blogs or micro blogs, uploading photos and connecting with other like-minded users. The part played by social media in biomedical field has emerged to be important in the past few years. Researchers and physicians have made use of the social media data to (1) to establish the communication and sharing of information between patients and health care decision holders, (2) design massive scale, dynamic disease surveillance systems and (3) mining of biomedical and health-related knowledge. Patients engage with health care organizations on social media with the anticipation that they will receive a friendly and helpful response. It has added plenty of opportunities for the patients in order to share their experiences with drugs and devices [1]. Pharmaceutical companies are preferring the monitoring of social network inside their IT departments, thus generating the possibility for the quick distribution and feedback on the products and services for optimizing and improving delivery, raise the turnover and profit, and limit the expenses.

Culotta proposed a methodology to study the predictability of Twitter data on future influenza rates [2]. A correlation of 95% was noticed between the tweets that contain the flu keywords and the real national health statistics. A similar study was conducted by Corley et al. who found a huge correlation between the frequency of the tweets (weekly) that contain the influenza keywords and the CDC influenza-like-illness surveillance information. Heaivilin et al. introduced Twitter as a potential source for dental surveillance and research [3]. The findings suggest that people who experience dental pain usually turn to social network to seek comfort and advice from others who also suffer from dental pain. In all such applications, systems are required to automatically, precisely, and effectively do the identification and interpretation of health-related information in compact text ‘‘micro’’ messages.

Although social media is soaring in noise owing to the heterogeneity observed of the styles of writing, formality, and creativity, such noise also bears undiscovered wisdom of the crowd, and hence should not be regarded as a threat, but an opportunity for discovering knowledge that can be useful in biomedical domains. Indeed literature illustrates rich research in mining biomedical and health related information in social media. Paul and Dredze utilized a modified Latent Dirichlet Allocation [4] model to identify 15 ailments along with descriptions and symptoms in Twitter data [5,6]. Yang10 applied association mining to analyze the relation between the drugs and their reaction. The drugs reaction was identified by detecting the keywords of the Consumer Health Vocabulary (CHV) in social media messages. These methods fail to capture the unknown dictionary words.

Extracting knowledge from social media becomes a most important to simultaneously improve healthcare which is analyzed and evaluated by the various researchers using different methodologies. In this research work, those research methodologies are discussed in terms of their working procedure and their functionalities along with various performance measures. The benefits and drawbacks that arise in those methodologies also discussed in the detailed manner.

2. EXISTING RESEARCH METHODOLOGIES
Corley et.al [7] introduced a new Text and Structural Data Mining method for Influenza Mentions in the Web and Social Media. Text and structural data mining of web and social media (WSM) yields a new disease surveillance resource and

can investigate about the online communities for the targeted public health communications (PHC) in order to guarantee an extensive distribution about the right kind of information. Text mining is proven to be identifying the trends in the flu posts, which have correlation to the actual-world influenza-like illness patient report information. This system brings to have a graph-based data mining method for detecting the deviations among the flu blogs linked by type of publisher, links, and user-tags. An introduced technique enables identifying the outbreaks and leads to an increase in the infection of influenza in the population.

Greene et.al [8] designed a new method for Qualitative Evaluation of Communication with Face book. Using the Face book search function, the proposed system searched for the word "diabetes" in the title of Face book groups. The system found the biggest groups on Face book that were focused on the patients affected with diabetes or people who are acquainted with or care for them. The findings show that the diabetes communities in Face book comprise of multifarious participants, inclusive of patients, family members, advertisers, and researchers, with different interests and ways of communication. These groups simultaneously act as promotional spaces, support communities, repositories with recruit able research subjects, and places for soliciting and provisioning forms of disease management-information that are generally not available through the more for malized means of professional consultation. The posts were then summarized and then gathered into a database. Two examiners assessed the posts, then designed a thematic coding methodology, and used the codes over the data.

Hirose et.al [9] introduced a new method for Predicting Infectious Disease Spread utilizing Twitter. An introduced methodology analyze the probability of developing a regression model by integrating the Twitter messages and Centers for Disease Control (CDC)'s Influenza-Like Illness (ILI) data, and it has been observed that the multiple linear regression model with ridge regularization performs better than the single linear regression model and other un-regularized least squared techniques. The model of multiple linear regressions with ridge can provide a notable improvement in the prediction accuracy.

Bodnar et.al [10] designed a Validating Models for Disease Detection Using Twitter. The multivariable regression is to perform regression on the keyword that correlates the best with the training data, and use it for the regression model. The system considers a form of regression that utilizes a SVM (support vector machine) which has been shown to predict ILI prevalence well. The designed system evaluated several well known regression models on their ability to accurately assess disease prevalence from tweets. It found that even irrelevant tweets and randomly generated datasets were able to assess disease levels comparatively well. This could serve as a ground level for evaluating other models: if a model can do only slightly better with seemingly relevant data than with seemingly irrelevant or random data, then it is probably not learning much from the tweets and its ability to fit the data can be attributed to other factors.

Cameron et.al [11] introduced a semantic web platform for drug abuse epidemiology making use of social media. The PREDOSE (PREscription Drug abuse Online Surveillance and Epidemiology) system first does the automation of the aggregation of web-based social media content for the next subsequent semantic annotation. The annotation scheme gets modeled in the DAO (Drug Abuse Ontology), and contains domain specific information like prescription (and corresponding) drugs, techniques of preparation, side effects, and means of administration. The DAO is also utilized in helping to identify the three kinds of data, which are: (1) entities, (2) relationships and (3) triples. Then, PREDOSE makes use of a combined lexical and semantic-based methodologies for the extraction of entities and relationships from the compiled content, and a top-down approach utilized for triple extraction, which in turn employs patterns that are expressed in the DAO. Moreover, PREDOSE makes use of openly available lexicons for identifying the initial sentiment expressions in text, and thereafter a probabilistic optimization algorithm (from related research) is used for extracting the final sentiment expressions. A past evaluation of the information extraction methods used in the PREDOSE platform shows 85% precision and 72% recall in entity identification, over a manually generated gold standard dataset.

Yang et.al [12] designed a new Social Media Mining method for Drug Safety Signal Detection. For the purpose of exploring the strength of identifying the Adverse Drug Reactions (ADRs) utilizing online healthcare communities, the system introduced the association mining and Proportional Reporting Ratios (PRR) for the extraction of amusing associations between drugs and adverse reactions. When the social media users contribute to the content concerned with the ADRs of a particular drug, the co-occurrence of the drug and it's ADR present in the posts or comments of an online healthcare social media site can be considered to be an association, and its attraction and impressiveness could be measured by examining such metrics to be support, confidence, leverage and lift. The FDA alerts are exploited to be the gold benchmark for testing the performance of the techniques proposed. Though all the three indicators have worked with efficiency in detecting the ADR based on the experiment, PRR and leverage have produced better results compared to lift.

Ji et.al [13] introduced a Twitter Sentiment Classifications for Monitoring Public Health Concerns. The designed system focuses on the sentiment classification of Twitter messages for measuring the Degree of Concern (DOC) of the Twitter users. To accomplish this objective, the system develops a new two-step sentiment classification workflow for automatically identifying the personal and negative tweets. Depending on this workflow, designed an Epidemic Sentiment Monitoring System (ESMOS), which renders the tools for visualizing the concern of the Twitter users' towards various diseases. The visual concern map and chart in ESMOS can assist the public health officials in identifying the advancement and peaks of concern for a disease in space and time, such that suitable preventive measures can be followed. The DOC measure is on the basis of the sentiment-based classifications. Finally, the clue-based and different Machine Learning techniques are compared for classifying the sentiments of Twitter users concerned with the diseases, initially into personal and neutral tweets and thereafter from neutral personal tweets into negative.

Tuarob et.al [14] introduced an ensemble heterogeneous classification technique for the discovery of health-related information in messages in social media. Its aim is addressing the drawbacks imposed by the conventional bag-of-word based techniques and suggests to make use of heterogeneous features combined with ensemble machine learning methods for discovering the health-related knowledge that could be helpful in several biomedical applications, particularly those requiring to find the health-related information in massive

scale social media knowledge. In addition, the methodology proposed can be generalized to find various kinds of information in different types of textual data. The system propose and test the efficacy of ensemble methods wherein multiple base classifiers that learn different facets of the data are utilized in combination to render collective decisions for enhancing the performance of health-related message grouping. Here two sets of evaluations are carried out in order to examine about the ability of the proposed model in discovering health-related data in the social media domain: small scale and massive scale evaluations.

Isah et.al [15] performed a Social Media Analysis for Product Safety employing Text Mining and Sentiment Analysis. An introduced system provides a report about a work in progress with contributions that includes: the design of a framework for the collection and analysis of the opinions and experiences of users about drug and cosmetic products utilizing machine learning, text mining and sentiment evaluation; the application of the newly introduced framework on the Face book comments and data from Twitter for the purpose of brand analysis, and the description about developing a product safety lexicon and training data for the modeling of a machine learning classifier (Naive Bayes) for drug and cosmetic product sentiment prediction. The initial brand and product comparison out comes prove the efficacy of text mining and sentiment analysis on the social media information whereas the usage of machine learning classifier for the prediction of the sentiment orientation yields a resourceful tool for the users, product manufacturers, regulatory and enforcement agencies for the monitoring of brand or product sentiment trends so as to act in case of an event of a sudden or considerable increase in negative sentiment.

Akay et.al [16] presents a Network-Based Modeling and Intelligent Data Mining of Social Media for the Improvement of Care. The system introduces a two-step analysis framework, which is focused on positive and negative sentiment, in addition to the side effects of treatment, in the forum posts of the users', and finds the user communities (modules) and influential users forgetting to know about the user views over cancer treatment. The system used a self-organizing map for analyzing the word frequency data obtained from the forum posts of users. Then a new network-based approach is introduced for modeling the forum interactions of the users' and used a network partitioning method on the basis of optimization of a stability quality measure. This allows to decide about the consumer views and find the influential users inside the retrieved modules making use of information obtained from both word-frequency data and network-based characteristics.

3. COMPARISON ANALYSIS

S.No	Reference	Method	Merits	Demerits	Results
1	Corley [7] , 2010	Graph-based data mining technique	It detects anomalies and informative substructures among the flu posts associated by publisher type, links, and user-tags.	It does not measure the effect and validate the usage of WSM for monitoring the seasonal influenza epidemics and global pandemics	The results signify unique WSM communities that are grouped by publisher and content type, like News Corp & Disney pr, international audiences, or personal blogs.
2	Greene [8], 2010	Qualitative Evaluation of Communication with Face book	It provides temporary support of the proposed public health advantages of social networking media in managing the chronic disease.	Inability to validate the identity of the poster— and the dominant usage of Face book pages for promoting non- FDA-approved therapeutic modalities — imposes a notable issue.	Face book yields a forum for reporting about personal experiences, posting questions, And getting a direct feedback for people who are suffering with diabetes.
3	Hirose [9], 2012	Multiple linear regression model	It improves the Prediction of Infectious Disease Spread accuracy	Better method needed for do the document classification work.	The model of multiple linear regression with ridge can significantly enhance the prediction accuracy
4	Bodnar [10] , 2013	SVM Classifier	Ability to accurately assess disease prevalence from tweets.	It cannot define diseases that have less predictable long term dynamics, such as gastroenteritis or asthma.	It may appear that both multiple regression and SVM regression have similar accuracies in the regional data.
5	Cameron [11], 2013	PREDOSE system	The PREDOSE platform shows 85% precision and 72% recall in entity identification.	It does not realize a module for entity disambiguation and improve the available	PREDOSE signifies 36% precision in relationship identification and

				modules for relationship, triple extraction and sentiment extraction	33% precision in triple extraction.
6	Yang [12], 2013	Social Data-based Prediction of Incidence and Trajectory model and collaborative prediction model	Collaborative Filtering (CF) is presented to predict a ranked list of future condition incidences.	Tree-based trajectory model needs to be improved	The framework is able to predict future conditions for online patients along with a coverage value of 48% and 75% for a corresponding top-20 and a top-100 ranked list.
7	Ji [13], 2013	ESMOS system	It facilitates the users to carry out the monitoring public health concern with time with a set consisting of visualization tools for epidemics-related Twitter information	Does not use the stream data mining methodologies for the constant modification and improvement of the model	Multinomial Naïve Bayes accomplishes the overall the best results and takes considerably lesser time to construct the classifier compared to the other techniques.
8	Tuarob [14], 2014	Ensemble machine learning schemes	The proposed system achieves better performance in terms of accuracy , precision and recall	To improve the classification algorithm and to employ semi-supervised methods such as the co-training technique to expand the training data with unlabeled data	Weighted Probability Averaging (WPA) method still offers considerably good performance with smaller performance degradation (17.88% drop in precision, 10.08% drop in recall, and 13.88% drop in F1)
9	Isah [15], 2014	Machine learning classifiers (Naive Bayes)	The method accurately groups the comments to be positive, neutral or negative	It does not consider temporal analysis for the detection of up or down trend of sentiment of a certain brand or product in addition to grouping the tweet and user sentiments by location.	The proposed technique achieves accurate result based on various factors like positive, neutral or negative comments.
10	Akay [16],2015	SOM technique	It is capable of investigating the positive and negative sentiment over the treatment of lung cancer by making use of the drug through the mapping of the large dimensional information onto a lower dimensional space employing the SOM.	It only concentrates on positive and negative opinions of the consumers which cannot give an accurate prediction of user opinions	Such an approach could be utilized to release red flags in future clinical surveillance operations; in addition to focusing over different other treatment related challenges.

4. INFERENCE OF EXISTING SYSTEM

In this study number of dataset such as TR-PN-s1w0 and gold standard dataset were taken. TR-PN-s1w0 comprises of 126,833 tweets, of which 63433 are personal tweets and 63400 are news tweets (8-fold re-sampled from 7,925 tweets). The gold standard dataset is created manually. The existing Graph-based data mining technique, Qualitative Evaluation of Communication with Face book, Multiple linear regression model, SVM Classifier and Social Data-based Prediction of Incidence and Trajectory model are used for discover the health-related knowledge from social media. But these methods are having major drawback in terms of constructing the medical ontology when the number of disease events to be observed, inability to verify the identification of the poster and it reduced the overall system performance by means of accuracy, precision and recall. In this survey, existing

PREDOSE, ESMOS, Ensemble machine learning, Naïve Bayes Classifier and proposed SOM methods are evaluated in terms of accuracy, precision and recall performance metrics. The proposed self-organizing maps (SOMs) method is utilized for assessing the correlations between the user posts and positive or negative opinion over the drug. And also it identifies the user communities (modules) and influential users accurately. The proposed system achieves better performance compared to other existing system while using various dataset.

3.1 Accuracy

It is defined to be the sum of the true positives and the true negatives, divided by the total number of classification parameters ($T_p + T_n + F_p + F_n$) .

$$\text{Accuracy} = \frac{Tp + Tn}{Tp + Tn + Fp + Fn}$$

Where,

Tp -True positive

Tn – Ture negative

Fp -False positive

Fn – False negative,

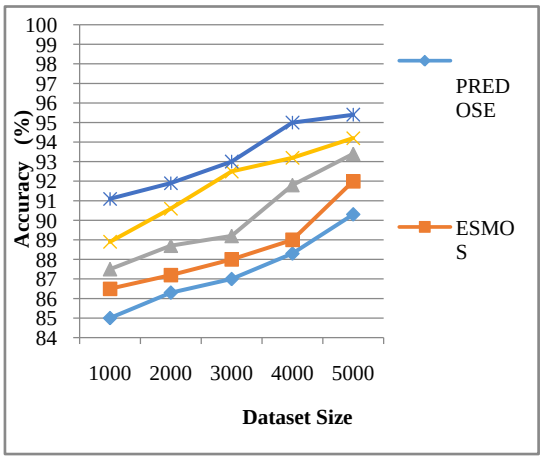

Fig1: Accuracy comparison

Figure 1 illustrates that the comparison of existing PREDOSE, ESMOS, Ensemble machine learning, Naïve Bayes Classifier and proposed SOM methods in terms of accuracy. For dataset size 5000, PREDOSE, ESMOS, Ensemble machine learning and Naïve Bayes classifier and SOM method achieves accuracy result of 91.1%, 91.9%, 93 %, 95 % and 95.4%. It concludes that the SOM method has shown the high accuracy results for all size of dataset.

3.2 Precision

Precision is defined to be the proportion of the true positives against both true positives and false positives results for intrusion and actual features. It is defined as follows

$$\text{Precision} = \frac{T_p}{T_p + F_p}$$

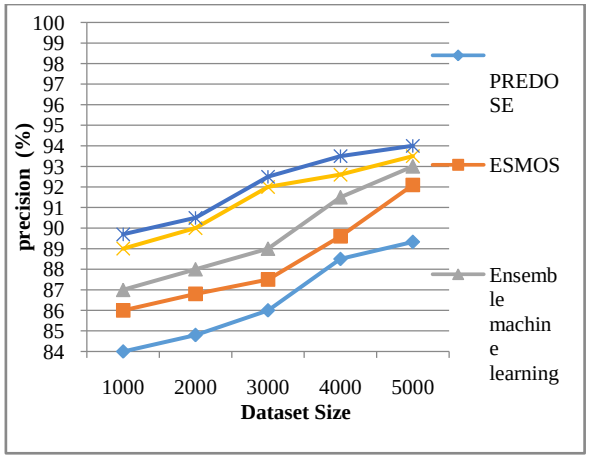

Fig 2: Precision comparison

Figure 2 shows the comparison result of existing PREDOSE, ESMOS, Ensemble machine learning, Naïve Bayes classifier and SOM method achieves precision result of 89.7%, 90.5%, 92.5 %, 93.5 % and 94 % respectively. From the graph it has been identified that the SOM method outperforms than that of the other models and results in precision values.

3.3 Recall

It measures the proportion of positives that are correctly identified

$$\text{Recall} = \frac{T_p}{T_p + F_n}$$

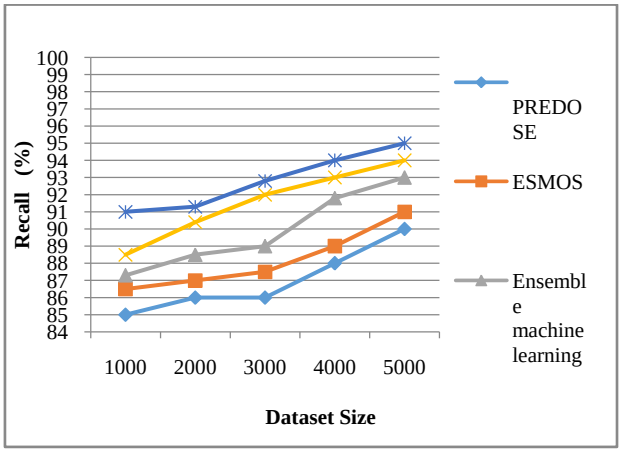

Fig 3: Recall comparison

Figure 3 illustrates that the comparison of existing PREDOSE, ESMOS, Ensemble machine learning, Naïve Bayes Classifier and proposed SOM method in terms of recall. For dataset size 5000, PREDOSE, ESMOS, Ensemble machine learning and Naïve Bayes classifier and SOM method achieves recall result of 91 %, 91.3%, 92.8 %, 94 % and 95 % respectively. It concludes that the SOM method has shown the high recall value for all size of dataset.

3.4 Overall Accuracy, Precision and Recall Comparison

The proposed SOM and existing PREDOSE, ESMOS, Ensemble machine learning, Naïve Bayes Classifier are evaluated in terms of Accuracy, Precision and Recall.

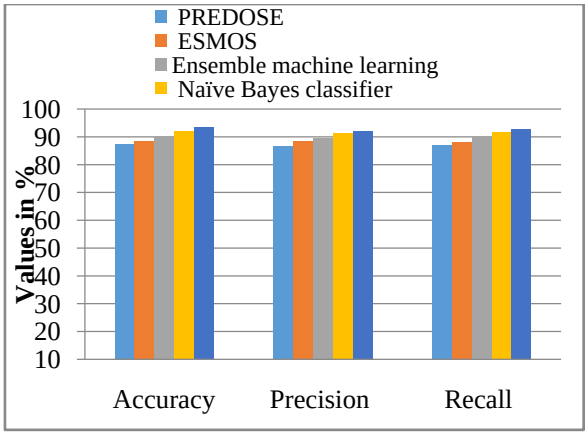

Fig 4: Performance Comparison

The SOM approach achieves recall as 92.82% which is 5%, 4 %, 2%, and 1% higher than PREDOSE, ESMOS, Ensemble machine learning, Naïve Bayes Classifier approaches respectively. From this graph, it concludes that SOM technique is efficiently discover the health-related knowledge from social media.

The newly introduced SOM methodology is compared with the existing PREDOSE, ESMOS, Ensemble machine learning, Naïve Bayes Classifier in terms of accuracy, precision and recall which is shown in Figure 4. It can be said that the proposed SOM approach has higher performance results when compared with the other existing PREDOSE, ESMOS, Ensemble machine learning, and Naïve Bayes Classifier methods. The proposed SOM provides the data to the network as a display, collecting together same kind of data weights to similar neurons. If new data is input into the network, the nearest weights that matches the data changes so that the new data gets reflected. The neurons which are at a distance from the new data changes rarely. Hence it is used to speedup of the overall process. Here it is observed that, the proposed SOM approach, the accuracy attained for all range of dataset size such as 1000,2000,3000,4000,5000 is 93.28%, which is 5.1%,3.4%, 2% and 1% higher than PREDOSE, ESMOS, Ensemble machine learning, Naïve Bayes Classifier approaches respectively. The proposed SOM approach achieves precision as 92.04% which is 4.67%, 1.9 %, 1%, and 0.5% higher than PREDOSE, ESMOS, Ensemble machine learning, Naïve Bayes Classifier approaches respectively.

5. CONCLUSION

The main intend of social media in biomedical knowledge mining, including clinical, medical and healthcare informatics, prescription drug abuse epidemiology and drug pharmacology, has rose to become remark ably important in the last few years. The various knowledge discovering framework was developed to predict risks of medical condition incidence and trajectories using patients' social media data. Those research methodologies are discussed along with their benefits and drawbacks in the detailed manner to find the effectiveness of every algorithm. The research works has been compared with each other based on their resultant metrics to find the better approach to precede the further research scenario in future. The final analysis of the research work tends to prove that the proposed SOM method is better under consideration of the all performance metrics which tends to provide the better result than the existing research methodologies.

6. REFERENCES

[1] Toldo L., 2013. “Text mining fundamentals for business analytics”, presented at the 11th Annual Text and Social Analytics Summit, Boston, MA, USA.

[2] Culotta, A., “Detecting influenza outbreaks by analyzing twitter messages”, CoRR abs/1007.4748.

[3] Heaivilin, N., Gerbert, B., Page, J and Gibbs, J., 2011. “Public health surveillance of dental pain via twitter”, J Dental Res., 90(9):1047–51.

[4] Blei, D.M., Ng, A.Y and Jordan, M.I., 2003. “Latent Dirichlet allocation”, J Mach Learn Res, 3:993–1022.

[5] Paul, M.J and Dredze M., 2011. “A model for mining public health topics from twitter”.

[6] Paul, M.J and Dredze, M, 2011. “You are what you tweet: analyzing Twitter for public health”, In: Fifth international AAAI conference on weblogs and and social media, 265–72.

[7] Corley and Courtney D, 2010. “Text and structural data mining of influenza mentions in web and social media”, International journal of environmental research and public health, 596-615.

[8] Greene, J., Choudhry, N., Kilabuk, E and Shrank, W, 2011. “Online social networking by patients with diabetes: a qualitative evaluation of communication with Facebook”, 26(3):287–92.

[9] Hideo Hirose and Liangliang Wan, “Prediction of Infectious Disease Spread using Twitter: A Case of Influenza”, Fifth International Symposium on Parallel Architectures, Algorithms and Programming, 2012.

[10] Bodnar, Todd and Marcel Salathé, “Validating models for disease detection using twitter”, In International Conference on Proceedings of the 22nd World Wide Web, 2013.

[11] Cameron and Delroy, 2013. “PREDOSE: A semantic web platform for drug abuse epidemiology using social media”, Journal of biomedical informatics, 985-997.

[12] Yang and Christopher C., 2012. “Social media mining for drug safety signal detection”, Proceedings of the 2012 international workshop on Smart health and wellbeing, ACM.

[13] Ji, Xiang, Soon Ae Chun and James Geller, 2013. “Monitoring public health concerns using Twitter sentiment classifications”, In International Conference Healthcare Informatics.

[14] Tuarob, S., Tucker, C.S., Salathe, M and Ram, N., 2014. “An ensemble heterogeneous classification methodology for discovering health-related knowledge in social media messages”, Journal of biomedical informatics, 49:255-268.

[15] Isah, H., Trundle, P and Neagu, D., 2014. “Social media analysis for product safety using text mining and sentiment analysis”, In 2014 14th UK Workshop on Computational Intelligence, 1-7.

[16] Akay, A., Dragomir, A and Erlandsson, B.E. 2015 , “Network-based modeling and intelligent data mining of social media for improving care”, IEEE journal of biomedical and health informatics, 19(1):210-21