IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

Automated Social Media Mining System in Health Care

Oleena Thomas

Assistant Professor, Department of Computer Science, Kottayam Institute of Technology and Science, Chengalam, Kottayam, Kerala, India

Abstract: Social media, by its nature, will bring different individuals with different experiences and viewpoints. Extracting knowledge from social media has great applications. In health care area, the advent in social media has created greater improvements in communicating. The users of social media post comments regarding different diseases and their remedies with the users' experiences in this regard. This could be so informative to other users as well. Others could get an overview regarding the diseases and their treatments. Taking the effect of social media into consideration, the information could reach a mass population. Sentiment analysis is the major focus here. Hence data preprocessing is a requisite. This is followed by the network modeling and side effect terms extraction. Through these the comments being considered for information extraction could be worth enough.

Keywords: Sentiment Analysis, Term Frequency, Inverse Document Frequency Scores, Data Pre-processing, Text Mining, Self Organizing Map.

I. INTRODUCTION

Data mining is highly sensitive to the data being dealt with. So the data collected must be preprocessed to maintain integrity between considered data. The reliability of the data from social media is guaranteed. Therefore, the data being mined is made to undergo preprocessing before reaching a conclusion.

Social media is one of the major modes of communication. It provides a better platform for the patients to discuss about their diseases as well as the drugs used for the same. This could be so informative to others as well. For instance, if a patient talks about a lung cancer, then he could discuss the drugs being consumed to overcome it as well as its side-effects. He could even suggest the best drug among many available in the market.

Considering the impact of social media, the discussion forums about health care could reach many people. Taking this into consideration, the discussion forums could act as source of data mining.
The valuable data can be mined out to reach a conclusion which could help in taking better decisions regarding the diseases and their treatments. For that, legitimate users must be identified. The discussion forums may contain influential users who could influence the innocent users. If there exists such users there are wide open chances of getting misled.

The networks modeling concept is used to model the social media users. The most active nodes are analyzed based on the modeling. The preprocessing involves the assignment of term frequency-inverse document frequency (tf-idf) scores. The words from the comments are considered relevant based on these scores. The side-effects of the treatment is also considered. The words qualifying such medical terms are also involved in comments. Medical terms are searched for in medical dictionaries.

Text mining has got extremely wide range of applications. Valuable information can be derived from text based contents using text mining. Text based contents may include word documents, emails and postings on social media. Nowadays text mining has got wide variety of applications in government, research, business etc. And these applications can be sorted into a number of categories based on the type of analysis and function. The major text mining tasks include text clustering, text classification, text categorization, sentiment analysis, document summarization, production of granular taxonomies etc.

Social media data should be made to undergo preprocess before some information could be retrieved. Sentiment analysis is one among the several techniques available. The polarity of each word could determine the orientation of the users' comments. This could be applied in health care sector which could be an aid to many patients, pharmaceuticals and the like.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

When sentiment analysis is applied to the preprocessed data, valuable information retrieval may take place. This is because, the polarity of the words towards positive, negative or neutral phases can be determined with sentiment analysis. Hence sentiment analysis is considered as an inevitable part in text mining.

To improve health-care outcomes and reduce costs using consumer-generated opinion, social media mining is an effective way. The web forums discussing health-care problems are to be taken into consideration so that from the user comments, useful information can be extracted. The positive as well as the negative sentiments from the comments are considered. The user communities are also identified. The influential users must be identified so as to ascertain the comments are from legitimate users rather than from misleading users. For the sake of the same network modeling is done.

The input social media comments are made to undergo preprocessing. The negative as well as positive sentiments are considered from the comments. This is followed by the identification of the real influential users so that the comments have the background of legitimacy. The comments may involve quotes regarding diseases, treatments, drugs and side effects. The words are given tf-idf score weightages. Based on the scores, the words are taken as information.

II. RELATED WORKS

A. Text Classification

With the proliferation of blogs, social networks and e- commerce sites, there is a great interest in supervised and semi-supervised text classification methods to reveal user sentiments and opinion to discover and classify the health service information obtained from the digital health ecosystems and to classify web resources for improving the quality of web searches. The problem of supervised text classification is that most times, a supervised classifier is unfeasible in a real context, because a large labeled training set is not always available. The labeling task for large datasets is practically unfeasible.

A complex vector of features is demonstrated in [3], based on weighted pairs of words, is capable of overcoming the limitations of simple structures when the number of labeled samples is small. Specifically, proposes a linear single label supervised classifier that is capable, based on a vector of features composed of weighted pairs of words, of achieving a better performance, in terms of accuracy, than existing methods when the size of the training set is about 1% of the original and composed of only positive examples. The proposed vector of features is automatically extracted from a set of documents D using a global method for term extraction based on the Latent Dirichlet Allocation implemented as the Probabilistic Topic.

Review selection [5] is another area where text mining methods are implemented. Review is concise i.e. up to 200 characters long and highly focused. A review is considered good, if it covers all of the attributes. The attributes are the different aspects of the entity. Both the positive and negative comments on multiple attributes are checked. Then the effective information is extracted from the review. This is done based on the greedy algorithm of effective coverage. In this the concept of tip is considered. For matching reviews and tips, semantic similarity is considered. A review sentence and a tip are syntactically similar if they share important keywords or if they share same concept or sentiment. For semantic similarity, the MALLET toolbox is used.

Top K high utility items can be mined using algorithms specified in [6]. But they are computationally ineffective as the user needs to specify parameters and multiple scans of database is a requisite.

Private FP growth tree can mine text by transaction split method [7]. Private FP growth (PFP-growth) algorithm consists of a preprocessing phase and a mining phase. In the preprocessing phase, the database is transformed to limit the length of transactions. The preprocessing phase is irrelevant to user specified thresholds and needs to be performed only once for a given database.
Several other semi supervised techniques [8] and [9] are used for text mining. Such approaches incur the cost of training the system which could prove ineffective.

B. Ranking Documents

With the quick development of Social Network Services (SNS) and Online Social Networks (OSN) based applications, people's information need is much more diversified and personalized so that the traditional search engine cannot always satisfy this new kind of information need. Consequently, they tend to seek information through their OSNs rather than the traditional search engine. Luckily, generating and sharing information is one of the most important properties of OSNs. It provides a natural way for people to communicate and go about seeking information. How to utilize users' social relation in the OSNs to improve search relevance is a challenging problem.

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

Users' activity degree and social influence [11] are applied as the ranking factors of the Social Relation Rank (SRR) algorithm. At the same time, a topic classification label is defined to adjust the weight of the algorithm. Topic belongs to different categories can be handled respectively according to its importance which leads to high computational cost.

Other methods were proposed to radically rank the influential users. In [4], initially the web forums are subjected to crawling and parsing. This is followed by data preprocessing. In the data preprocessing, metadata extraction is done initially. Then follows cleaning process. User lists, quotations and body of the text are separated. User radicalness identification is done by analyzing the user posts against the threat list. Then the radicalness is identified with the help of collocation matrix. Based on the computation of associativity radically influential users are ranked. As in the other ranking system, computational efficiency is poor.

C. Community Detection

Many real systems can be represented as networks whose analysis can be very informative regarding the original system's organization. In various fields such as biology, sociology, engineering and beyond, systems are commonly represented as graphs, or networks. Communities are defined as groups of nodes that are more densely connected internally than with the rest of the network. They can be of unequal size and density. The quality of a community structure can be assessed using several criteria. The most common and explored quality function is modularity as defined in [10].

According to [12], online social network consists of a number of users that among them there are social interactions. A social network has a social graph and log of actions. By running the frequent pattern mining algorithm, the maximal patterns are obtained from user-action table. This is used in frequent pattern mining. Well-known frequent pattern mining algorithms like Apriori, Fp-Growth can be used. Each maximal pattern makes a community, if its nodes, regarding threshold are related to each other.

In network dynamics, other methods like artificial societies, social emulation [2] using algorithms like cultural algorithms, ant colony, particle swarm optimization are involved. As the others, these methods also incur high computational cost.

III. METHODS

A forum focused on oncology is converted into weighted vectors to measure consumer thoughts on the drugs used using positive and negative terms alongside another list containing the side effects. The methods are able to investigate positive and negative sentiment on lung cancer treatment using the drug by mapping the large dimensional data onto a lower dimensional space using the SOM. Most of the user data is clustered to the area of the map linked to positive sentiment, thus reflecting the general positive view of the users. Subsequent network based modeling of the forum yielded interesting insights on the underlying information exchange among users.
A two-step analysis framework that focuses on positive and negative sentiment, as well as the side effects of treatment, in users' forum posts is proposed for ascertaining user opinion of cancer treatment. A self-organizing map is used to analyze word frequency data derived from users' forum posts. Then a novel network-based approach is introduced for modeling users' forum interactions with four different types of nodes.

A. Initial Data Search and Collection

Most popular cancer message boards are searched and out of them lung cancer has been focused as per the statistics lung cancer is the most widespread among cancer. A list of drugs used by lung cancer patients was complied to ascertain which drug was the most discussed in the forums. The drug Erlotinib (trade name Tarceva) was the most frequently discussed drug in the message boards. A further search revealed that Cancerforums.net, despite having slightly fewer posts on lung cancer, had more posts dedicated to Erlotinib than the others.

B. Initial Text Mining and Preprocessing

A data collection and processing mechanism is developed to look for the most common positive and negative words and their term-frequency-inverse document frequency (TF-IDF) scores within each post. The fig. 1 shows the diagrammatic representation of preprocessing [1].

The user posts are given as input in the data preprocessing section. The data is read in and then is made to undergo some filtrations so as to obtain a valid set of data. The filtration may include content extraction, tokenization, case transformations and filtering of stop words and tokens. Thus the documents are converted to data. This eliminated excess noise (misspelled words, common stop words, etc.) to ensure a uniform set of variables that can be measured. Processed data contains the final word list, with each word containing a specific tf-idf score.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

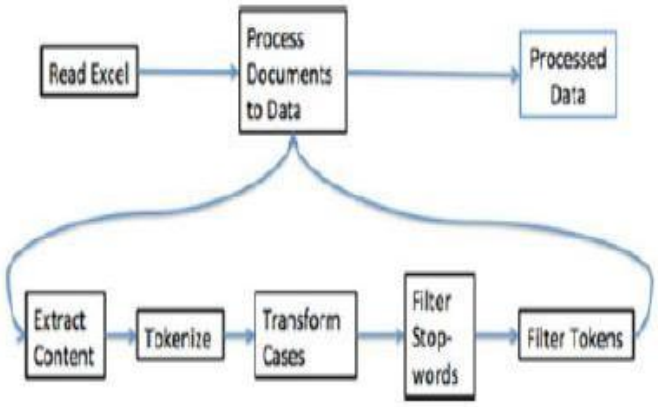

Fig. 1 Processing mechanism

C. Cataloguing and Tagging Data

Text data containing the highest TF-IDF scores are tagged with a modified NLTK toolkit to ensure that they reflected the negativity of a negative word and the positivity of a positive word in context.

This approach was used before using negative tags on positive words. A positive tag is added on negative words. The NLTK toolkit is then used for the analysis, and classification, of words to match their exact meanings within the contextual settings.

Words should be marked positive or negative with respect to the contextual meaning so as to reflect the correct orientation of the comments. For instance, consider the comment “I do not feel great". In this, “great" may indicate positivity. But the comment is negative. Now consider another statement, “No side effects so I am happy". Here the word “no" tends to be negative but in the statement context, it is positive. Words which are not frequent are eliminated. Thus a statistically significant set of data is obtained. The end list resulted in a set of positive and negative words.

In parallel to this, the side effect terms are also extracted. Such terms were searched for in the medical databases. All the side effect terms satisfying threshold tf-idf scores are then grouped to form a list. Thus after the preprocessing technique, two word lists are obtained: one contains list of positive and negative words while the other contains the list of side effect terms.

D. Using a Self Organizing Map

SOMs are neural networks that produce low-dimensional representation of high dimensional data. Within this network, a layer represents output space with each neuron assigned a specific weight. The weight values reflect on the cluster content.

The SOM displays the data to the network, bringing together similar data weights to similar neurons. The existence of clusters in the data and how the SOM weights of these clusters are assessed would correlate to positive and negative opinion.

When new data is fed into the network, the closest weights matching the data change to reflect the new data. The neurons farther from the new data rarely change. This process continues until data is no longer fed, resulting in a two-dimensional map.

The SOM toolbox is used and the SOM is fed with first word list TF-IDF vectors. The purpose is to assess the existence of clusters in the data and how the SOM weights of these clusters would correlate to positive and negative opinion. The word list data is mapped and the weights are analyzed for positive and negative variable correlations of the word list. Words of no interest, and groups containing three or fewer words, are then eliminated.

E. Modelling Forum Postings using Network Analysis

Networks were built from forum posts and their replies, while accounting for content-based grouping of posts resulting from the existing forum threads. Based on the posts, the different users are modeled as nodes of four different types namely transmitter, isolated, carrier and receptor. Optimal information modules could be taken into consideration for further analysis.

The TF-IDF scores from the word list of positive and negative words can be used to build the node types. TF-IDF scores to the average of the collected posts of the user is examined to obtain the local measure that illustrates specific user opinion to each node in the module.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

F. Network based Identification of Side-effects

The TF-IDF scores within each module will directly reflect how frequent a certain side-effect is mentioned in module posts. Then the next step is to compare the values of the TF-IDF scores within the module to those of the overall forum population and identify variables (side-effects) that have significantly higher scores.

IV. EXPERIMENT RESULTS

The performances of the natural language tool kit as well as the performance of the Self Organizing Maps are witnessed. Positive and negative word lists along with list of side effect terms are generated by the natural language tool kit. Initially, the SOM is fed with the first word list generated from tf-idf score calculations. Then SOM performs the data clustering so as to cluster the similar data. The weight values reflect on the cluster content. When new data is fed into the network, the closest weights matching the data change to reflect the new data.

The neurons farther from the new data rarely change. This process continues until data is no longer fed, resulting in a two-dimensional map.

A. Initial Data Search and Collection

The data was collected from the forum “cancerforums.net". All the posts under the topic of lung cancer were collected.

B. Classification Accuracy

Using the SOM classifier, the correct classification rate is high when compared to the other classifiers like REP tree and Naive Bayesian classification.

C. Observed Data

The SOM classifier classifies with maximum accuracy. But with respect to the time considered, SOM is less accurate than Naive Bayesian classifier but is accurate than the REP classifier. The figure 2. shows the classifiers' efficiency with respect to time. The accuracy is 76.67% using SOM classifier whereas it is 90% for Naïve Bayesian classifier at the expense of time and computational cost and 40% for REP tree classifier.

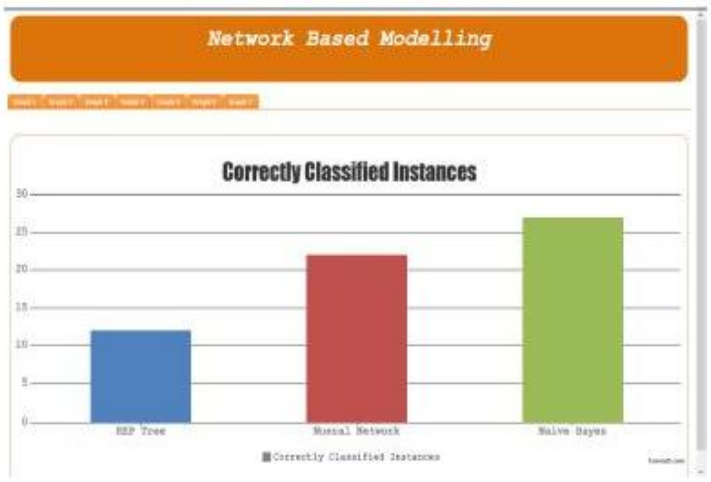

Fig. 2 Correct classification of instances

V. CONCLUSIONS

Extracting knowledge from social media has recently attracted great interest from the Bio-medical and Health Informatics community to simultaneously improve health care outcomes and reduce costs using consumer-generated opinion. 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.

Pharmaceutical companies are prioritizing social network monitoring within their IT departments, creating an opportunity for rapid dissemination and feedback of products and services to optimize and enhance delivery, increase turnover and profit, and reduce costs. Social media data harvesting for bio-surveillance have also been reported.

Web forum is converted into weighted vectors to measure consumer thoughts on the drugs, diseases as well as side-effects. Both the positive and negative sentiments were considered by mapping the large dimensional data onto a lower

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

dimensional space. Network based modeling of the forum yielded interesting insights on the underlying information exchange among users. Modules of strongly interacting users were identified as influential users.

Identification of potential side-effects could alert the clinical surveillance operations, as well as highlighting various other treatment related issues. This solution can be envisioned on future medical devices that can serve as post-marketing feedback loop that consumers can use to express their satisfaction or dissatisfaction directly to the company. The company benefits from real-time feedback that can then be used to assess if there are any problems and rapidly address such problems.

REFERENCES

[1] Altug Akay, Andrei Dragomir and Bjorn-Erik Erlandsson, "Network Based Modeling and Intelligent Data Mining of Social Media for Improving Care," IEEE Transactions On Biomedical and Health Informatics, vol. 19, no. 1, Jan. 2015.
[2] Alberto Ochoa, Aturo Hernandez, Laura Cruz, Julio Ponce, Fernando Montes, Liang Li and Lenka Janacek, "Artificial Societies and Social Simulation using Ant Colony, Particle Swarm Optimization and Cultural Algorithms," Proc. Int. New Achievements in Evolutionary Computation Feb. 2010.
[3] Francesco Colace, Massimo De Santo, Luca Greco and Paolo Napoletano, "Text classification using a few labeled examples," Computers in Human Behavior, Aug. 2013.
[4] Tarique Anwar and Muhammad Abulaish, "Ranking Radically Influential Web Forum Users," IEEE Transactions On Information Forensics and Security, vol. 10, no. 6 Jun. 2014.
[5] Thanh-Son Nguyen, Hady W. Lauw and Panayiotis Tsaparas, "Review Selection Using Micro-Reviews," IEEE Transactions On Knowledge and Data Engineering, vol. 9, no. 4, Apr. 2014.
[6] Wen Zhang, Xijin Tang and Taketoshi Yoshida, "TESC: An approach to Text classification using Semi-supervised Clustering," Knowledge-Based Systems, Nov. 2014.
[7] Vincent S. Tseng, Cheng-Wei Wu, Philippe Fournier-Viger, and Philip S. Yu, "Efficient Algorithms for Mining Top-K High Utility Itemsets," IEEE Transactions On Knowledge and Data Engineering, vol. 8, no. 1, Jan. 2014.
[8] Sen Su, Shengzhi Xu, Xiang Cheng, Zhengyi Li and Fangchun Yang, "Differentially Private Frequent Itemset Mining via Transaction Splitting," IEEE Transactions On Knowledge and Data Engineering, vol. 8, no. 1, Jan. 2014.
[9] Mohammad Salim Ahmed and Latifur Khan, "SISC: A Text Classification Approach Using Semi Supervised Subspace Clustering," IEEE International Conference on Data Mining Workshops, 2009.
[10] Erwan Le Martelot and Chris Hankin, "Multi-Scale Community Detection using Stability Optimization," Int. J. of Web Based Communities, 2012.
[11] Liang Guo, Xirong Que, Yidong Cui, Wendong Wang, Shiduan Cheng, "A Hybrid Social Search Model Based on the Users Online Social Networks," Proceedings of IEEE CCIS, 2012.
[12] Seyed Ahmad Moosavi and Mehrdad Jalali, "Community Detection in Online Social Networks Using Actions of Users," IEEE International Conference on Data Mining Workshops, 2014.

BIOGRAPHY

Oleena Thomas, working currently as Assistant Professor CSE in Kottayam Institute of Technology and Science, Kottayam has pursued her M. Tech in CSE from Amal Jyothi College of Engineering Kottayam and B. Tech from College of Engineering Kottarakkara. She has published papers named "Literature Analysis on Reputation Models for Feedback in E-commerce" and "Data Mining Approach to Wind Data Preprocessing" in IJARCCE.