ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

Intelligent Data Mining from Healthcare Forums

M.Rajeswari

M.E Student, Dept of C.S.E, Arunai Engineering College, Thiruvannamalai, India.

ABSTRACT: A huge amount of data which is generated by the user is available in the medical forums with the help of online health care social media. This social media helps the patients to play an important role in the medical forums like get the information about medicine, treatment. The patients can actively interacting with other patients or physicians in the online healthcare discussions. But they can't find the related topics in that medical forum. It is some what not easy to the users. Because the data's in medical forums are in unstructured format. Recommendation systems are based on content based approach which recommends about discussions to the patients. But the content based approach is not sufficient. Because the patients may have different intentions or they are having interest in different types even if the data matches their interest. By using Naïve Byes method, there are two tasks are proposed for classifying posts and comments in the medical forums. They are Classification of intentions and Classification of social support types. Inorder to develop the Classifiers, text feature sets and health feature sets are used. To combine Classifiers with different feature sets and also optimizing the classification results, the genetic algorithm is used. For Post Classification, combining text and health feature sets can achieve the highest precision, recall and F1 measure. For Comment Classification, combining the text feature sets can achieve the best result.

KEYWORDS: Classification and Optimization Techniques,Genetic Algorithm, Naive Bayes Algorithm,Recommender Systems.

I. INTRODUCTION

Data mining, the withdrawal of secreted prognostic data from large databases, is a powerful new technology with huge possible to aid companies focus on the most important information in their data warehouses. The process of ascertain pattern in large data sets relating method at the assembly end of artificial intelligence, machine learning, and shared database systems etc. The main goal of the data mining process is to dig out the information from a data set and transform it into an clear arrangement for prospect utilize. The large datasets generated from databases are being mined to extract secreted knowledge that are useful for decision makers to take effective, efficient and timely decisions in a competitive world. Data withdrawal tools expect prospect trends and behaviors, allow business to create practical, knowledge-driven decision. The automated, potential analysis accessible by data withdrawal move beyond the analysis of past actions provide by display tools distinctive of resolution hold systems. Data withdrawal tools can reply business question that usually were too time overwhelming to decide. They clean databases for secreted pattern, finding predictive data that expert may miss because it lies exterior their prospect. Many companies previously gather and process huge quantity of data. Data withdrawal techniques can be implement quickly on obtainable software and hardware platforms to develop the rate of obtainable information capital, and can be included with original products and system as they are bring on-line. When implemented on high recital client/server or parallel processing computers, data withdrawal tools can examine huge databases to distribute answers to questions. This white paper provide an opening to the essential technologies of data withdrawal Examples of gainful applications demonstrate its significance to today's business situation as well as a fundamental explanation of how data warehouse architectures can develop to distribute the value of data withdrawal to last part users.

A large number of online communities and groups are developed for people to talk about health problems with attributes of social networking, involvement, apomediation, group effort and directness. A huge amount of customer developed substance can be found taking place healthcare social media, which provides information and source to examine user attributes and events. QuitStop is a trendy forum of smoking termination in QuitNet website. Registered QuitNet customers can involve in interactions and communicate with each other on QuitStop forum. They communicate tobacco stopping process, raising questions and receive social support.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

Diverse ways that computer science investigate helps solving healthcare issues. One of the significant ways is empower users to participate a important task in their own health and treatment in replicated surroundings. Health communities on diverse social media, likewise QuitStop forum, have the possible to provide such utility, because users can be totally concerned in the interactions ,which is apart from time and locations. However, owing to the huge content and unstructured content of information on QuitStop forum or various health communities, it is complex for patients to find related topics or peers to interact.. It would be useful if recommending the attractive topics or calculate possible patients for QuitStop forum. Recommendation techniques have been useful in online forum for topic recommendation and customer predictions. Collaborative filtering and substance based approaches are two main approaches. However, it is exposed that collaborative filtering suffers from cold create issues. Because customer relations are thin in online forums. For QuitStop forum or various healthcare communities, recommendation straightly on text substance may not execute well, due to the small communications and the use of diverse vocabularies in relating healthcare problems. In healthcare social media, health users may have diverse objectives of involvement or may be attracted in diverse types of social support. In addition to, the health condition of health users may also point toward their particular interests.
It is frequent ,that text attributes are used for categorizations of online communications. For a thread on QuitStop forum, text could be mined from title, post and comments. Choosing the various text attributes to develop classifier individually, and group them with specific weights to minimize the categorization effect. Furthermore, mining the health data from patient profiles, and use their health condition as attributes to develop text categorization. Researches are considered to decide the attributes that realize the most excellent presentation.

II. RELATED WORK

A. Online Communities of Health involvement:

A bunch of online health communities have been industrial on diverse social media sites. Various studies apply qualitative analysis to examine the content of user chat and build up diverse classification schemes . Our earlier research paying attention on user communications of social support replace and extracted five themes from communication on QuitStop forum, as well as present social support, requesting social support, getting social support, other actions and unrelated content.. Societal support is “a return of property connecting two individuals supposed by the contributor or the receiver to be proposed to improve the security of the receiver”. It is significant for health involvement programs to facilitate patients in establishing encouraging thoughts and assurance. In online health communities, social support is exchange linking diverse users. Chuang and Yang extract two major type of social support from thought of an online alcoholism group of people, which are informational support and nurturant support.Informational support is also called exploit facilitate support and task-oriented support. It is given to facilitate patients answer or remove health troubles. Nururant support, projected by Cutrona and Suhr is also called socio-emotional support. It is given to relieve and calm patients without straight hard work to solve troubles. In studies of smoking cessation, quit status is an significant health quality of smoking quitters. It also reflect the result of smoking cessation involvement. For users of QuitStop forum, some of their quit dates could be extract from their profile pages, from which the quit position could be considered.

In our earlier investigate the quit position of a user is considered by the number of days that he/she has been temperate from the self-reported day he/she stop smoking on the profile page to the day he/she post the communication. It takes a extended point for people to quit smoking. According to quit statuses, users of QuitStop forum could be separated into five quit stages. Users with the quit statuses of 0 to 3 months are at Stage 1 – premature exploit stage; users with quit statuses of 3 to 6 months are at Stage 2 – delayed exploit stage;Users at Stage 3, premature continuance period, are those who have been quitted for 6 months to 2 years; Those with quit statuses of 2 years to 5 years are at Stage 4 – delayed continuation period; and those who have been temperate for more than 5 years are at Stage 5, which means that they have finished smoking cessation. In this study, quit status and quit stage are regard as health features, and they are used for categorization.

B. categorization of User Generated substance:

In online forums, inquiring and answering are the common events and user communications that could be extracted from conversation content. For inquiry answer detections, dissimilar categorization methods are used to examine user generated content. Kim et al classify threads in a student conversation board. They measured speech perform patterns and projected methods to identify conversation focal point. However, their categorization is implemented physically on a tiny dataset. Antonelli and Sapino industrial a rule-based classifier to recognize the family of diverse postings.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

In some studies , diverse categorization techniques are useful and compared for inquiry/answer recognition, with algorithms of maximum entropy, SVM and CRF. A diversity of features were also preferred and compared. categorization method is also used to appraise the quality of post content. To examine the fullness, solvedness, spam and trouble types of threads in a Linux user forum, Baldwin et al extracted text features from diverse positions of threads, and used diverse categorization algorithms for thread categorization.

In health part, categorization and other information mining techniques are useful to examine structured biological data, such as attributes of cells, genes, proteins, etc. Academic articles in Medline/PubMed are confidential based on text attributes. For social media examination of healthcare, text mining and social network examination are used to spread infectious diseases with hospital report, predict pandemic increase with Twitter data, model hospital arrangement network, or examine health social network for a few websites. Some studies useful text mining techniques to examine the post content in online health forums and groups. However, these studies did not examine the full text of user generated substance.

III. SYSTEM ARCHITECTURE

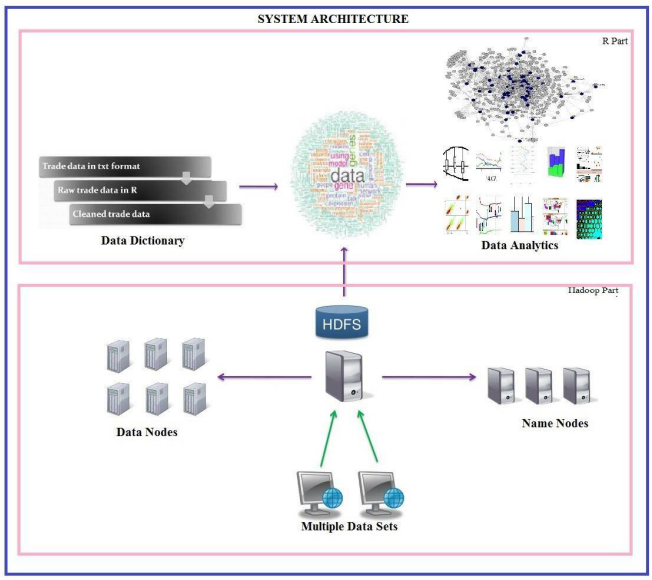

Figure 1: System Architecture

A. Task explanation:

In this study, we categorize posts and comments based on diverse features. Two tasks are projected for post categorization and comment categorization, respectively: categorization of user intentions, and categorization of social support types. With qualitative examination in our earlier study,five themes were extracted from communication of QuitStop forum, which are contribution social support, requesting social support, receiving social support, other activities and unrelated substance. These themes reveal user intentions to issue matching posts or comments. These five categories are used for the categorization of user intention in this study. For categorization of social support types, two categories are urbanized, which are informational support and nurturant support. According to earlier definitions, informational support is exact information about the disease, treatment or coping, as well as subcategories of advice, referral, fact, perceptual information, personal experiences and feedback/view.Nurturant support is express caring or concern, and express the significance of association, as well as esteem, network and emotional support. These two types of social support are not restricted. A message could be assign to equally informational support and nurturant support.

B. Feature explanation:

In this study, we construct classifiers with diverse types of feature sets, and join diverse classifiers to get the top results. Two types of feature sets are used for categorization: text quality sets and health 5quality sets of users.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

1) Text quality Sets:

Generally, message categorization is based on text features. On QuitStop forum, a thread is consisted of texts at diverse positions. We build classifiers with text features at diverse thread positions, i.e. title, post, and comments, and join different features linearly. Title refers to the title of the equivalent thread that is formed by the post author. Post refers to substance of the post in the related thread formed by the post author.

2) Health quality Sets:

A few users of QuitStop forum rendering their quit date in their profile pages, from which we can compute their quit statuses and quit stages. Our earlier research establish that there are assured relations between user quit status and their actions to print diverse types of communication. So, in this study, we make use of quit status and quit stage as health to recover text classifications.For the author of each communication in our datasets, we symbolize his/her quit status by the number of days that he/she has been ascetic from the self-reported day he/she stop smoking on the profile page to the day he/she posts the communication. According to quit statuses, users of QuitStop forum could be separated into five quit stages as mentioned. Based on quit status and quit stage, diverse health quality sets are industrial as shown in Table 1. PA status, PA stage, CA status and CA stage are the four health quality sets. But the facts of their meanings differ in post categorization and comment categorization.

3) *Valuation Metrics:*

For a categorization task, precision, recall and F1 score could sbe intended for each class. They can assess the result of classifications. For a convinced class, let n be the number of reports that are predicted in the class, N denotes the number of reports belong to the class in land truth, and tp be the number of reports that are predicted in the class correctly, the precision could be intended as P=tp/n, the recall could be intended as R=tp/N, and the F1 score is intended as F1=(2*P*R)/(P+R).

Health quality Sets	Explanation in Post categorization	Explanation in Comment categorization
PA Status	Quit status of the post author	Quit status of the post author in related thread
PA Stage	Quit stage of the post author	Quit stage of the post author in related thread
CA Status	Mean value of the quit statuses of all comment authors in related thread	Quit status of the comment author
CA Stage	Quit stages of all comment authors in related thread	Quit stage of the comment author

Table 1:Health quality Sets

IV. EXPERIMENTAL RESULTS

A. Classification of Posts:

1) Classification Task of Intentions:

Center on the classes of contribution social support and request social support in this job for post categorization. All the precisions, recalls and F1 scores report in this job are the average standards of that of the two classes. Expand classifiers with text attribute sets of title, post and comment independently. The classifier with title attribute set reach the highest precision, recall and F1 score. Then, join diverse text attribute sets, and make use of health attribute sets to increase the categorizations. Researches are carried over with goals to enhance precision, recall and F1 score, correspondingly.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

2) Classification Task of Social Support Types:

Likewise,develop classifiers on texts of title, post and comment independently. The precision, and F1 score can reach the maximum value with post text. The classifier develop on title has the maximum recall, With goals of optimizing precision, recall and F1 score correspondingly, a sequence of researches are carried over to join diverse text attributes and health attributes. Enchanting precision as the optimization aim, the result of researches. The classifier only with post text has the maximum precision. Addition of various text attributes or health attributes could not very much develop the precision.

B. Classification of Comments:

1) Classification Task of Intention:

For the objective categorization of comments, concern the classes of request social support and in receipt of social support. All the precisions, recalls and F1 scores report are considered on these two classes. Started from increasing classifiers with text attribute sets of title, post and comment correspondingly. The classifier by means of post attribute has the maximum precision and F1 score correspondingly. The classifier with title attribute set reach the maximum recall.

Analyze the mixture weights of diverse text attribute sets, it could be fulfilled that the post substance is significant to point toward the objective of comments. Addition of comment substance is useful to develop precision, while adding up title can facilitate to develop recall and F1 score.

2) Classification Task of Social Support Types:

Text-only classifiers are develop on attribute sets of title, post and comment correspondingly. The classifier with title attribute has the maximum precision and the classifier by means of comment attribute reach the maximum recall and F1 score of correspondingly. Observing text classifiers that reaches maximum precision, recall and F1 score, suggested that comment substance is significant for the categorization.Addition of title with a maximum weight could very much ccdevelop recall as well as F1 score, but it cannot improve precision.

V. CONCLUSION AND FUTURE WORK

In this study, we relate categorization and optimization techniques to classify posts and comments on QuitStop forum. Diverse text quality sets and health quality sets are used to construct classifiers. The categorization outcome are analyzed and compared. Shortening all experiments in this study, it is experiential that:

(1) For a thread on QuitStop forum, the substance formed by the post author (title and post text) can point to the meaning of both post and comments in that thread. For intention tasks
of both post and comment classifications, the optimization procedure assigns high weights to the quality sets of title and post.

(2) For the categorization of social support types, the communication substance is an significant quality set. Concretely, post substance is valuable for post categorization, and comment substance is valuable for comment categorization.

(3) Health quality sets are valuable to progress post classifications of both intentions and social support types. However, none of the health features sets can progress comment classifications considerably. User generated substance on healthcare social media provides a large quantity of information to point to user features and communications. In this study, we concern categorization and optimization techniques to recognize user intentions and social support types from a forum of smoking cessation involvement, which is a addition of our earlier works. In the prospect, we shall expand recommender systems for topic recommendation and user predication based on our categorization outcome. Additional techniques and experiments would be conducted to examine user generated substance of online healthcare social media.

REFERENCES

1. Ming Yang,Melody Kiang “Extracting consumer health expressions of drug safety from web forum” 2015 48th Hawaii International conference on system sciences.
2. Taridzo Chomutare “Patient similarity using network structure properties in online communities” 978-1-4799-2131-7/14/$31.00 ©2014 IEEE.
3. Juan Li,Nazia Zaman “Personalized healthcare recommender based on social media”2014 IEEE 28th International Conference on Advanced Information Networking and Applications.
4. Priya Nambisan, ZhihuiLuo, Akshat Kapoor,Timothy B Patrick “Social media,big data and public healthcare informatics:ruminating behavior of depression revealed through twitter” 2015 48th Hawaii International Conference on System Sciences.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

5. Jasmine Bhaskar,Sruthi K,Prema Nedungadi "Enhanced sentiment analysis of informal textual communication in social media by considering objective words and intensifiers" IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2014), May 09-11, 2014, Jaipur, India.
6. Asha Menon,Fallon Farmer "Automatic identification of alcohol-related promotions on twitter and prediction of promotion spread" IEEE 2014.
7. Karolos Talvis,Kostantinos Chorianopoulos "Real-time monitoring of flu epidemics through linguistic and statistical analysis of twitter messages" IEEE 2014.
8. Tatsuhiro Sakai,Keiichi Tamura and Hajime Kitakami "Density-based adaptive spatial clustering algorithm for identifying local high-density areas in georeferenced documents" IEEE 2014
9. Hong Qing Yu,Xia Zhao,Xin Zhen "Healthcare-event driven semantic knowledge extraction with hybrid data repository" IEEE 2014.
10. Haruna Isah,Paul Trundle,DanielNeagu "Social media analysis for product safety using text mining and sentiment analysis" IEEE 2014.