Intelligent mining for Disease Prediction using Short Texts

Mrs. Rakhi Akhare[1], Dr.S.D.Sawarkar[2]

1 (Asst.Prof. Computer Engineering Dept L.T.C.E, Navi Mumbai, India)
2 (Principal, Datta Meghe College of Engineering, Navi Mumbai, India)

ABSTRACT: Machine Learning is used to make a computer system intelligent to give exact and updated output to users of that system. So it becomes as an important technology almost in all domains particularly in research and medical fields. This proposed system avoids the unnecessary information from web search results like advertisements and gives results as per user's convenience. It also provides updated knowledge which deals with current research and discoveries in medical domain. Machine Learning is done by user's personal experiences or from web databases like MEDLINE. This system can be used by lay people as well as experts, doctors to improve their knowledge related to various diseases and treatments as it is up to date with current researches. This paper deals with various task sets and data representation techniques which can give improved results. This system can be used in different medical care system to take better medical decisions. It can save time of people for searching medical data and avoids confusion as it gives results in short texts.

Keywords: EHR, Machine Learning, Medline, Natural Language Processing, UML.

I. INTRODUCTION: The traditional healthcare system is also becoming one that hug the Internet and the electronic world. Electronic Health Records (EHR) is becoming the standard in the healthcare domain. Researches and studies show that the potential benefits of having an EHR system are as Health information recording and clinical data repositories immediate access to patient diagnoses, allergies, and lab test results that enable better and time-efficient medical decisions; Medication management rapid access to information regarding potential adverse drug reactions, immunizations, supplies, etc. Decision support the ability to capture and use quality medical data for decisions in the workflow of healthcare; and Obtain treatments that are tailored to specific health needs—rapid access to information that is focused on certain topics. Already existing approaches mostly lack on automatic information mapping and they struggle to extract particular information from a clustered form of data. They were far behind in classification performance. Using of search engines results in drawbacks such as poor precision, poor recall, varied document quality and in varied indexing path. The main drawback of using search engines involves a learning curve. Many beginning internet users because of these disadvantages become discouraged and frustrated. The tasks that are addressed here are the foundation of an information technology framework that identifies and disseminates healthcare information. The framework's capabilities can be used in a commercial recommender system and it is integrated in a new Electronic Health Record system.

II. RELATED WORK: People care deeply about their health and want to be, now more than ever, in charge of their health and healthcare. Life is more hectic than has ever been; the medicine that is practiced today is an (Evidence-Based Medicine hereafter, EBM) in which medical expertise is not only based on years of practice but on the latest discoveries as well. Tools that can help us manage and better keep track of our health such as Google Health1 and Microsoft HealthVault2 are reasons and facts that make people more powerful when it comes to healthcare knowledge and management. The traditional healthcare system is also becoming one that embraces the Internet and the electronic world. The most relevant related work is the work done by Oana Frunza, Diana Inkpen, and Thomas Tran [1]. Their work performed two tasks in pipelined manner for identifying and extracting the relationship between the given MEDLINE abstract. First task involves finding most suitable model for prediction; the second task is to find good data representation. To achieve this two task various predictive algorithm and textual representation techniques are considered. A set of six classification algorithm namely decision based models, probabilistic models(Naïve Bayes, Complement Naïve Bayes), Adaptive learning, linear classifier namely support vector machine and a classifier that always predicts the majority class in training data are used. The advantages and limitations of the entire six classification algorithm are discussed. Three representation technique namely Bag-Of-Word representation, NLP and Biomedical

Concept representation and Medical concept representation are used to obtain the treatment relation from short text. Various experiments are conducted with the combination of the six classification algorithm and three representation techniques. The results are shown in bar chart form. As the result of the experiment it is concluded that bag-of-representation when combined with any of six classification algorithm produces better results. Limitation occurs when the remaining two representation technique is used.

Rosario et al. [2] introduced the Machine Learning (ML) Approach for Identifying Disease-Treatment Relations in Short Texts The main focus of their work is on entity recognition for diseases and treatments. The authors use Hidden Markov models and maximum entropy models to perform both the task of entity recognition and the relation discrimination Their representation techniques are based on words in context, part of speech information, phrases, and a medical lexical ontology—Mesh6 terms. The system contains both informative and non informative sentences. So the fast access of reliable information is not possible. This is the major drawback of the existing system. Also it is difficult to classify the sentences because of the eight semantic relations which also lead to confusion. The task of relation extraction or relation identification is previously tackled in the medical literature, but with a focus on biomedical tasks: sub cellular location (Craven, [3]), gene-disorder association (Ray and Craven, [4]).

MEDLINE is the largest component of PubMed, the freely accessible online database of biomedical journal citations and abstracts created by the U.S. National Library of Medicine. Approximately 5,400 journals published in the United States and more than 80 other countries have been selected and are currently indexed for MEDLINE. PubMed Comprises more that 22 million biomedical literatures from Medline, life science journals and online books which can be available here (http://www.ncbi.nlm.nih.gov/pubmed). Since 1990, the MEDLINE database has grown faster than before with more documents available in electronic form. The cost of human indexing of the biomedical literature is high, so many attempts have been made in order to provide automatic indexing. A unique feature of MEDLINE is that the records are indexed with NLM's controlled vocabulary i.e. the Medical Subject Headings (MeSH). Medical Subject Headings (MeSH) mainly consists of the controlled vocabulary and a MeSH Tree. (www.nlm.nih.gov./mesh). MeSH descriptors have been used to index PubMed articles and used as features to extract information from PubMed articles [5] used MeSH descriptors as the selected features for classification and showed that there is a significant improvement of classification performance.

III. PROPOSED SYSTEM: In the proposed work user will search for the disease summary (disease and treatment related information) by giving symptoms as a query in the search engine. In this work, two databases are used. One is Medline database and the other is Local database. The Medline database contains more than 21 million records from approximately 5,000 selected publications covering biomedicine and health from 1950 to the present. MEDLINE uses Medical Subject Headings (MeSH) for information retrieval. Approximately 5,000 biomedical journals are indexed in MEDLINE. The local database contains user's uploaded information related to different diseases and treatments. User may be any lay people, doctor etc.

In this examine, we focus on diseases and treatment information, and the relation that exists between these two entities. The proposed system focused on two tasks. The first tasks will automatically identifying sentences published in medical abstracts as containing or not information about diseases and treatments and also automatically identify semantic relations that exist between diseases and treatments as expressed in these texts. The second task is focused on three semantic relations: Cure, Prevent, and Side Effect. It uses pipelining means one task is followed by other i.e. output of one task becomes the input of other. It removes additional efforts to classify irrelevant data. We can get correct results in less number of steps.

Tasks Set:

Registration: In this task, users such as Doctors, Patients, Hospital Staff, Medical experts want to upload the diseases details and treatment details. For that purpose they first registered his/her personal details and then login into page to upload medical files.

Upload Files: Here, we describe about that the users are upload the medical details like disease details, with the treatment details, etc., it is very useful for the many people. They can easily retrieve the data from the upload database. Each disease have separate id to generate. It stores the data in the database.

Input processing: The task is performing the input get the some word of medical sentence from uploaded files or through the web database. The word contains matching word to search from the web domains. The rules are used to determine if a textual input contains relations. The sentences in which the relation appears and the local context of the entities. Before begin to extract Disease-Treatment relation about a particular disease from Medline, save a particular page describing the information of the disease from Medline as a .html file and store in a desired location or in a specific database. The system architecture of the proposed system describing the methodology of this project is below The html file containing user mentioned disease is saved with .html extension to the user specified storage namely a database or in research repositories. The next step is to convert the file with .html extension as .txt extension. This is input for further processing.

Web Medical Database: The web domain storing medical web databases like Medline. The Medline database contains more than 21 million records from approximately 5,000 selected publications covering biomedicine and health from 1950 to the present. MEDLINE uses Medical Subject Headings (MeSH) for information retrieval. Approximately 5,000 biomedical journals are indexed in MEDLINE. The data response the domain, then searching and reply data output to the user level. The domain of research and just recently has become a reliable tool in the medical domain. The experimental domain of automatic learning is used in tasks such as medical decision support, medical imaging, protein-protein interaction, extraction of medical knowledge, and for overall patient management care.

Training: The task is performing the input get the some word of medical sentence from uploaded files or through the web database. Since the system is designed to self learning capability, it should be trained and feature set should be collected from the training data set. Training Dataset may be user's uploaded file or Web database (Medline) abstracts. The training contains two tasks. Firstly, it identifies sentences which contains user relevant information and divide into two parts, relevant and irrelevant data. Secondly it identifies that sentences contains any semantic relation out of three relations i.e. cure, prevent and side effect.

This involves removing all the HTML tags, frames, images, ordered, unordered list, cascading style sheets and it retrieves, stores only the text content in the html file as text file with .txt extension. The obtained text file may be stored in any of the location mentioned by the user. Here after the content in the text file will be processed by using various classification algorithms, association rules and representation technique to achieve a processed text file which contains only the information related to Symptoms, Causes, and Treatment about the disease in the user specified .html document.

Note that all the process must be followed in pipelined manner in order to achieve a high quality result. Now the extracted text file contains many stop words like a, an, is, for, of, words ending with ing, ed etc. These words can be removed to improve the quality of the result. Thus stop words are removed from the extracted text file. Now the stop word removed text file is subjected to the combination of certain words in order to avoid repetition such as excessed, excessing if both these words appear in the document we shall reduce the word count just by using stemming algorithm which results word as excess removing the suffixes like ed, ing. It is very common that all the documents will contains such repetition of words for user to get clear understanding of sentences or information. By removing such suffixes and combining these kinds of sentences the content of the document is reduced but the quality of the document is increased by reducing the word count and describing the information in simpler form. After applying stemming algorithm, the semantic relations should be extracted from the above processed text file. Here the semantic relation is the information related to Symptoms, Causes and Treatment of certain disease in the user uploaded html file. After that keywords which are included in users query needs to match with processed text file. It gives semantic relation between keywords and information contains in that file. We will get some information which is useful for user. Now it is important part to present that information to users in proper format.

Classification: After sentences and relations identification, it is required to classify those sentences in relevant classes and represent results in such format which is easily understandable by any user of this system. For this following methods are used:

In order to extract this semantic relations a classification algorithm namely Discriminative Multinomial (DMNB) Naïve Bayes classification algorithm is used in association with Apriori association rule mining. Naïve Bayes specially used for text documents. NB models the word count and performs the classification within it. Apriori association mining is used to find co-occurrence of features in the form of association rules. Textual representation technique for labeling the training data and for identifying the sentences related to the label Symptoms, Causes and Treatment are achieved by using Bag-Of-Word representation technique along with word sequence pattern is used.

Bag-of-Words Representation: The bag-of-words (BOW) representation is commonly used for text classification tasks. It is a representation in which features are chosen among the words that are present in the training data. Selection techniques are used in order to identify the most suitable words as features. After the feature space is identified, each training and test instance is mapped to this feature representation by giving values to each feature for a certain instance.

Concept Dictionary Representation: After identifying semantic relations and feature extraction it is require to assign a proper naming representation which is more general than words in abstract. UMLS is a knowledge source developed at the US National Library of Medicine and it contains a met thesaurus, a semantic network, and the specialist lexicon for biomedical domain. The met thesaurus is organized around concepts and meanings; it links alternative names and views of the same concept and identifies useful relationships between different concepts. It is just like dictionary in which one word can have different meanings.

Search Engine: The output performs the must be need of the exact web medical data and relational data's get to here. We extracted only noun-phrases, verb-phrases, and biomedical concepts as potential features from the output of each sentence present in the data set. Output of this proposed method should give correct and relevant results to user's search. Here user searches for any queries related to their health and they will get exact answer in very small text which is very simple to understand.

Output Performance: Now the resulted file containing information related to Symptoms, Causes, and Treatment from the uploaded html file is tested for its quality. The quality of the resulted file is obtained by calculating its Precision, Recall, F-measure.
The formulas for calculating these quality measures are Precision= (relevant + retrieved) document/ Retrieved document. Recall = (relevant + retrieved) document/ Relevant document. F-measure= mean of Precision and Recall. A bar chart is used to represent the amount of word counts in the resultant file with their precision. Recall and F-measure value. The above performed Disease-Treatment information classification and extraction can be used in applications like medical domain, Online patient information storage system, Research scholars and Doctors, Patients to update their knowledge in a particular domain and in bio-informatics.

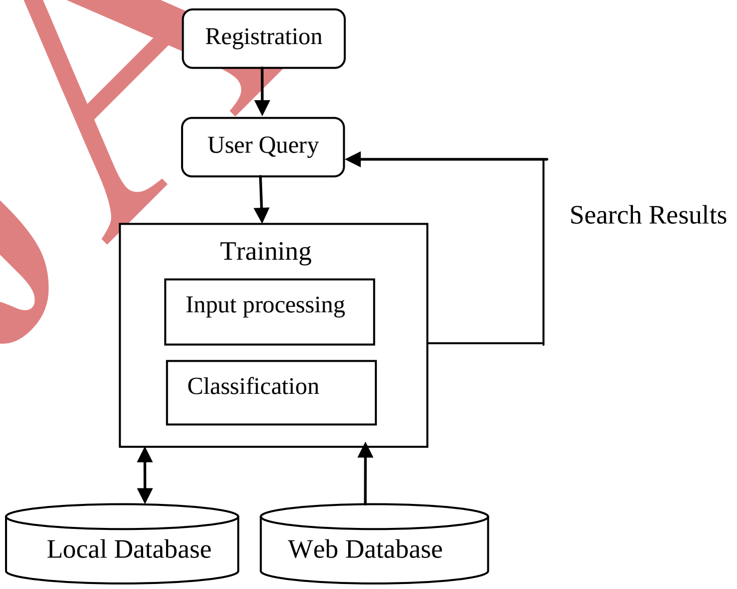

fig 1. System Architecture of Proposed System

As shown in Fig.1, after registration user can upload his/her file which is stored in local database of system. Along with this user again can search for details of particular disease-treatment information. Related to that search, information can be extracted from local database or from web database. After information retrieval, it is processed with various techniques as explained above. Finally relevant results will be provided to user in simple language using short texts. The proposed system gives up-to-date information as it extracts data from web which stores all current research and discoveries in medical domain. It saves time and efforts of user which is the main drawback of existing system.

IV. CONCLUSION: As per various studies and experiments, this paper will give improved results in medical domain. Machine Learning approach eliminates the need of physicians to manage user databases. This system again deals with latest research and discoveries in medical domain, so no need to update database manually. Firstly, according to user's query, it retrieves the information by matching keywords then processes that information to give relevant results to user. It uses different representation techniques which can very easily understandable by user. The doctor can easily update their knowledge and gain idea about particular medicine that are effective for some patient but causes side effect to patient with some additional medical disorder. It can again save time and efforts of people while searching healthcare information and avoid confusion as results appear to them in short texts.

Future Enhancement is to use various web medical databases to retrieve information, as here we used only one web database i.e. MEDLINE. Some more classification and representation techniques can also be added in processing of data like some images, videos etc.

REFERENCES

[1] Oana Frunza, Diana Inkpen, and Thomas Tran "A Machine Learning Approach for Identifying Disease-Treatment Relations in Short Texts", Member, *IEEE-VOL. 23, NO. 6, JUNE 2011 801*.

[2] B. Rosario and M.A. Hearst, "Semantic Relations in Bioscience Text," *Proc. 42nd Ann. Meeting on Assoc. for Computational Linguistics, vol. 430, 2004.*

[3] M.Craven, "Learning to Extract Relations from Medline," *Proc. Assoc. for the Advancement of Artificial Intelligence, 1999.*

[4] Suchitra A. and Sudha R**.,** "Extraction of Semantic Biomedical Relations from Medline Abstracts using Machine Learning Approach" *IJCA, NCACSA 2012.*

[5] P.Bhaskar, 2E.Madhusudhana Reddy "Efficient Machine Learning Approach for identifying Disease-Treatment Semantic Relations from Bio-Medical Sentences" *,IJCER Vol. 2 Issue. 5.*

[6] M. Goadrich, L. Oliphant, and J. Shavlik, "Learning Ensembles of First-Order Clauses for Recall-Precision Curves: A Case Study in Biomedical Information Extraction," *Proc. 14th Int'l Conf. Inductive Logic Programming, 2004.*

[7] http://en.wikipedia.org/wiki/Microsoft_HealthVault.

[8] http://en.wikipedia.org/wiki/Google_Health.

[9] "*Data Mining: Practical Machine Learning Tools and Techniques"* Third Edition (The Morgan Kaufmann Series in Data Management Systems)

[10] "*Data Mining: Concepts and Techniques*" Second Edition (The Morgan Kaufmann Series in Data Management Systems.

[11] Elmasary and Navathe "*Database Management System*" Fifth Edition, Vikas publication House