International Journal of Computer Applications Technology and Research
Volume 4– Issue 3, 184 - 187, 2015, ISSN:- 2319–8656

Identifying Disease-Treatment Relation Using ML and NLP Approach

Dhamne Kalpesh
SIEM, Nashik,
University of
Pune, India

Mistari Sachin
SIEM, Nashik,
University of
Pune ,India

Dahite Sushil
SIEM, Nashik,
University of
Pune, India

Dalvi Suraj
SIEM, Nashik,
University of
Pune, India

R. S. Shirsath
SIEM, Nashik,
University of
Pune, India

Abstract: This paper presents the efficient machine learning algorithm and techniques used in extracting disease and treatment related sentences from short text published in medical papers. **.** In this paper better machine learning algorithms and techniques are used for extracting disease treatment relations from various medical related articles. The proposed system gives the user exactly the Disease and Treatment related sentences by avoiding unnecessary information, advertisements from the medical web page namely Medline. For making better medical decisions we can make use of this proposed technique.

Keywords: Machine Learning, Disease Treatment, Medline, Stemming Algorithm & Natural Language Processing Multinomial Naive Bayes algorithm

1. INTRODUCTION

Now a day's people are more aware about their health and healthcare. In spite of their busy schedules they want information regarding to their health for each and everything in a suitable way. People want Fast access to reliable information and in a manner that is suitable to their habits and work-flow. Medical field has grown in a wider to such an extent that information about latest discoveries are published day by day. The proposed system gives more reliable information and classification performances regarding medline database. Our proposed technique provides the doctors in making better medical decisions.

Medline is the database which contains the latest medical articles with disease and treatment information. Medical related article are very large. Now days to read complete articles published in these databases is not possible. It is a tedious work. So to avoid such problems we extract informative sentences related to disease, treatment, and three semantic relations between them like cure, prevent and side effect [3].

Our proposed system is to work with NLP and ML technique which has the task of identifying and disseminating information. The work that we present in this paper is focused on two tasks:

Task1: It automatically identifying sentences published in medical abstracts (Medline) containing information about diseases and treatments, and identifying semantic relations that exist between diseases and treatments. Task1 is done by using the "stemming algorithm".

Task2: It is focused on three semantic relations: Cure, Prevent, and Side Effect. This project will be more useful for common users who find difficulty in reading medline. It will be done by using "Multi nominal Naive Bayes algorithm" [1].

2. LITERATURE SURVEY

The work presents various Machine Learning (ML) and information for classification of short texts and finds the relation between diseases and treatments. As per the ML technique related information are shown in short texts when identifying relations between two entities such as diseases and treatment. It is better to identify and remove the sentence that does not contain information relevant to disease or treatments [4]. The remaining sentences can be classified according to the relation. It will be very difficult to identify the exact solution if everything is done in one step by classifying sentences based on interest and also including the sentences that do not provide relevant information. The data set used in this work is UMLS. The data set contains information from medline with all relevant information including diseases, treatments and eight relations between diseases and treatment. For mapping the words into semantic categories they used medical subject headings. In this work they compare different graphical models and generative models. For extracting semantic relations the Naive Bayes algorithm is used [3].

Although this system does not provide accurate results these systems are successful in this biomedical field. New rules have to be followed each time because the semantic rule based system has a disadvantage that lexicon changes from domain to domain. To get the good results semantic and syntactic based systems are combined so that they provide flexibility of syntactic information and good precision of semantic rule. Statistical approaches are used to solve different tasks. So that rules will extract automatically [2]. This method is used to solve different NLP tasks. So this approach works well even with fewer amounts of data. Considering relation extraction the rule checks whether the text information contains any relation or not. The statistical approach uses bag-of-words technique for the relation extraction. Some researchers combined this technique with POS which provides two sources of information such as relation between their specific contexts and entities. Since it is proved that simple technique can produce accurate results [4].

The traditional healthcare system is also becoming one that hug the Internet and the electronic world. In the healthcare domain, Electronic Health Records (EHR) is becoming the standard. Studies and researches show that the potential benefits of having an EHR system are:

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; The EHR is web base application require server client communication it is very critical to maintain the connection

for long term use [3]. These disadvantages overcome in propose system.

In this system individual sentences are considered as instances that are to be processed by the naive bayes classifier. Here each instance is considered as positive training set. Alternative relation extractions are made through relational learning. Relational learning involves parsing sentences and from the parsed sentences, parse tree is constructed. From the parsed tree grouping of the relevant sentence made. The extracted results are in proper form. The task of relation extraction was previously tackled in medical literature for gene-disorder association. It involves automatic extraction of relation between medical concepts. UMLS is used for finding the medical concept in sentence classification. Using semantic parser the sentences are automatically parsed. After applying semantic extraction a set of extraction, alteration, validation rules are applied to distinguish the actual semantic relation to be extracted [2].

3. PROPOSED SYSTEM

In this proposed system for easily identifying and collecting the healthcare information's published in various medical related midlines. The difficult problem here is that to know about a particular disease and its treatment people have to read the entire article. So in order to avoid such tedious work we provide them with an easy method of extracting only related or informative sentences from the medical articles. So here people get the information regarding a particular disease in the form of three semantic relations cure, prevent and side effects. We also find the symptoms focused in the articles related to disease. For removing the unwanted information from the articles we use many methods [1]. We drop out the stop words form the articles and then by using the stemming algorithm we remove the repetition of words and after that with the help of Multinominal Naive Bayes algorithm and semantic probability calculations extract the informative words. The application used is designed using dot net. The command named relation finder finds the relation between diseases and treatments and also provides us other information's. Whenever the button is pressed the user or doctor obtains the relevant information regarding that particular disease. In order to improve the quality of the result the process are performed in a sequential manner. To avoid uninformative sentences we first perform the stop word removal. We remove stop words such as a, an, is, any, about, of, if, in etc. from the text file. There are about 174 English stop words and we remove the entire stop words from the text file so that we can improve the quality of the result. By stop word removal content is reduced but quality is improved to a greater extend [4].

Next step is removal of repeated words from midline. We know that after the stop word removal process the remaining text file contains repeated words such as expressing and expressed etc. The stream of such words for example express is same for two words we combine both of them to one word so that the repetition can be avoided. all the repeated words are removed. This removal of repeated words will increase the quality of result to a much upper level. For the removing the repeated word we use the suffix stemming algorithm. There are many different stemming algorithms that we are known. From this different stemming algorithm here we use the suffix stripping algorithm [1].

We have to find the disease treatment relations from the remaining text document. In the form of three semantic relations cure, prevent and side effect. We also find the symptoms associated with the disease. For finding the semantic relations here the Multinominal Naive Bayes algorithm is used. The algorithms will easily finds the relation and we can easily display it to the end user. Naive Bayes algorithms drawbacks are overcome in Multinominal Naive Bayes.

In text classification we make use of this Multinominal Naive Bayes algorithm due to its computational advantage and simplicity. The algorithm is a specialized version of Naive Bayes [3]. The Naive Bayes algorithm is not used here because it suffers from some drawbacks. The major difference is that it assumes that the attributes of a given class are not dependent on each other. In some cases the attributes are related to each other. For example consider the classifier for in the case of assessing the risk of issuing a check book. For a worthy customer it will not be true to assume that there is no dependency or relation between that customers age, worth and education status. We prefer Multinominal Naive Bayes algorithm to avoid this problem. In naïve Bayes algorithm we calculate the semantic probability, which helps in easily recognizing the disease treatment relation.

The above described method of finding disease treatment relation can be used in various other applications in future work. The result quality can be found out with the help of f-measure values, recall and precision [1]. This will helps in saving the time of various users especially doctors by easily extracting the informative sentences from the medical related midline. There are various important modules used to perform these task and they are described as follows

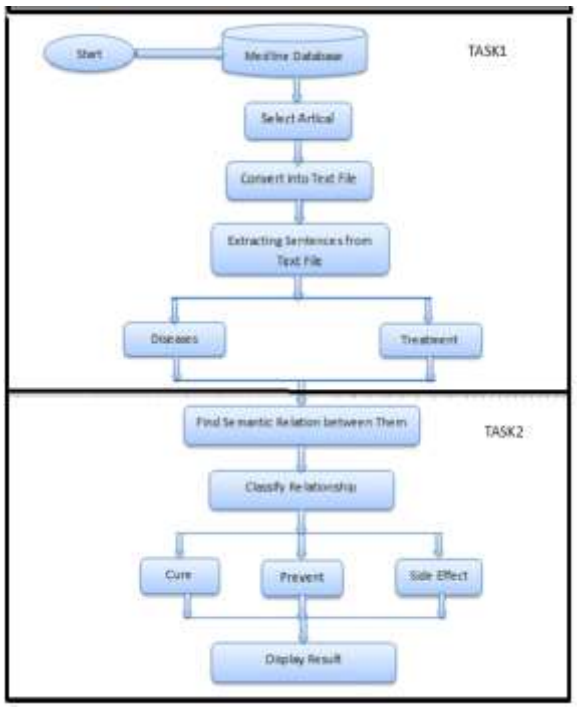

Fig 1: System Architecture

3.1 Html to text conversion:

The saved .html document is converted into a text file and is stored with .txt extension. This involves removing all the HTML tags, 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 location mentioned by the user [1].

3.2 Extraction of informative data:

Bag-Of-Word (BOW) representation is used for text classification where each of the word is used as feature for training the classifier in training dataset. BOW represents a document as a frequency of word occurrences. This classification and representation is unable to maintain any sequential information. In our proposed system, Weighted Bag-Of-Word representation is used to overcome the drawbacks of above mentioned problem of BOW [1].

3.2.1 Stop Word Removal Process:

As the first process we remove the stop words associated with each sentence. After the stop word removing the content size is reduces & document quality is improves. There are about 174 English stop words and all these when present in the document are successfully removed. Ex. a, an, is, for, the etc.

3.2.2 Repeated words Removing:

After the removal of stop words the remaining document contains repeated words and phrases and these words have to be removed from the contents extracted from above to improve the quality of the contents . To remove the repeated words and phrases we use the stemming algorithm. But the stemming algorithms has different types. Out of this algorithms here we make use of the suffix stripping algorithm. This may be done by removal of the various suffixes like -ED, -S, -ING, -ION, -IONS. For Ex.-

GENERALIZATIONS
GENERALIZATION
GENERALIZE
GENERAL

3.2.4 *Sentence Identification And Relationship Extraction:*

From the extracted contents, related with a disease and its treatment the three semantic relations such as cure, prevent and side effects are find out. To resolve the above problem and to result in efficient sentence identification Multi-nominal Naive Bayes classification algorithm is used in the proposed system. This algorithm is mostly used for the text classification. This algorithm finds the relations between disease and treatment and we can easily display it to the user by using related data set. Multi-nominal Naive Bayes classification (MNB) algorithm adopts parameter learning method [4].

3.2.4 *Output Performance Evaluation:*

This proposed system output is evaluated for various medline abstracts. The results we obtained shows informative sentences relevant to disease, treatments and the three disease treatment relations and symptoms related to the disease. The different data sets are used to extracting information associated to the three semantic relations that are cure, prevent and side effects. The predictable model is created to show the information regarding above mentioned semantic relations [3].

4. EVALUATION AND RESULT

The performance measurement is the efficiency of solution to given problem. It considers the performance of the trained models which yields the best predictive and classified results from the test dataset. Various standard measures gives the better score in relation extrication which is relevant to our problem domain. Ex. Accuracy which is measured by, Accuracy = total corrected corrections /total predictive [3].

From the recovered sentences, choose a testing dataset and a training dataset. ML setting worked on the training dataset and computed against the testing dataset. It gone very simple for selecting randomly in separation of data (Ex. 63% in training dataset, 37% in testing dataset) or may contains more complex sampling or extrication methods. But while processing on both datasets, they should be represent the solution for the problem.

4.1 Evaluation & performance Measures:

The important evaluation measures in ML algorithms are: accuracy, recall, precision and F-measure. As per the predictive concept of a model: confusion matrix (figure out the accuracy, cost of classification, F-measure). We can calculate ROC curve, and roles of every classifier is shown as a point on ROC curve. Whenever changes in the threshold value in the algorithms, cost matrix of classification, the point locations on ROC curve will alter respectively [1].

All above mentioned measures are evaluated to form a confusion matrix which includes information of the true classes, the actual classes and the classes prophecies by classifiers. The test dataset on which the predictive models are calculated include the true classes and the performance tries to recognize how many of true classes were forecasted by the model classifier. In the ML algorithms, focus needs towards the evaluation or performance measures that are used [4].

4.2 Efficiency of Identifying Informative Sentences:

This gives the evaluation for the first task, i.e. sentences are positive or negative (informative or non-informative). The ML algorithms are predicted for classification and represented as described above. Results of a classifier give the majority for improvement of datasets [7].

4.3 Efficiency of Identifying Semantic Relations:

Second task recognises sentences which contain information about 3 semantic relations like Cure, Prevent, and Side Effects. While performing operations on imbalanced data, F-measure is reported [3].

4.4 Performance of overall system:

In second task solution, to find the semantic relation we compare the results in 4 classes: 3 semantic relations and set of non-informative sentences. Performance of overall system can be computed as an evaluation measures of first task (results of classifiers) and second task (reporting F-measure results for imbalanced data).

4.5 Future Work:

These predictive models have stability and reliability for various tasks brings off on short texts in the medical field. The classification techniques gives more impact on ML algorithm results, but more informative classified results are the ones that regularly gives the best results to the users. The first task fulfilled in this paper is a task that has applications in recovery of important information, extricate the recovered information and text categorization. When more information is available for extrication or classification, there is an improvement in forecast results.

BOW method yields the best results in the text summarization or information extrication, can be more relevant when attaching more information from different kinds of things. The second task that performed can be seen as a task that could give profit by performing the first task first. To perform a handling or sorting of the sentences to get results for a relation classification. We adjoined the information from relation extraction that includes any of the three relations like disease, treatment and preventions, and excluded the sentences which did not contain above three

semantic relations. This search is very useful in over out the positive and negative sentences before classification or extrication of those sentences.

5. CONCLUSION

It provides reliable & efficient medical information in short-text. The proposed work provides us only informative sentences and removes uninformative sentences from the medical related articles in a pipelined manner. This system helps users especially doctors in saving their time and they can know easily about a disease its treatment and symptoms and can analyses more about a various treatments associated with a particular disease. This system will be more useful to common users who want to know more about a disease in simpler manner.

6. REFERANCES

[1] Oana Frunza, Diana Inkpen, and Thomas Tran, Member "A Machine Learning Approach for Identifying Disease-Treatment Relations in Short Texts" june 2011.

[2] Ancy Sudhakar and Merin Meleet "A System for Extraction of Semantic Biomedical Relations Using Multinominal Naive Bayes Algorithm", March 2014.

[3] Janani.R.M.S and Ramesh V.," Efficient Extraction of Medical Relations using Machine Learning Approach", March 2013.

[4] Mouratis, S.Kotsiantis, "Increasing The Accuracy Of Discriminative Of Multinominal Bayesian Classifier In Text Classification**",** ICCIT'09 Proceedings Of The 2009 Fourth International Conference On Computer Science And Convergence Information Technology.

[5] R. Kohavi and F. Provost, "Glossary of Terms," Machine Learning, Editorial for the Special Issue on Applications of Machine Learning and the Knowledge Discovery Process, vol. 30, pp. 271-274, 1998.

[6] J. Li, Z. Zhang, X. Li, and H. Chen, "Kernel-Based Learning for Biomedical Relation Extraction," J. Am. Soc. Information Science and Technology, vol. 59, no. 5, pp. 756-769, 2008.

[7] T.K. Jenssen, A. Laegreid, J. Komorowski, and E. Hovig, "A Literature Network of Human Genes for High-Throughput Analysis of Gene Expression," Nature Genetics, vol. 28, no. 1, pp. 21-28, 2001