International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 4 | Issue 2 | ISSN : 2456-3307

Review On Machine Learning Approach for Detecting Disease-Treatment Relations in Short Texts

Alapati. Janardhana Rao*[1], Reddy Srinivasa Rao[2]
[1*] janardhan182@gmail.com, srinukunta78@gmail.com[2]

ABSTRACT

The Machine Learning (ML) field has gained its momentum in almost any domain of research and just recently has become a reliable tool in the medical domain. Empirical domain of automatic learning is used in tasks such as medical decision support, protein-protein interaction, medical imaging, and extraction of medical knowledge. ML is envisioned as a tool by which computer-based systems can be integrated in the healthcare field in order to get a better and more efficient medical care ML-based methodology for building an application that is capable of identifying and disseminating healthcare in-formation. Due to advancements in medical domain automatic learning has gained popularity in the fields of medical decision support, complete health management and extraction of medical knowledge. The main objective of this work is to show what Natural Language Processing (NLP) and Machine Learning (ML) techniques used for representation of information and what classification algorithms are suitable for identifying and classifying relevant medical information in short texts. This paper describes how ML and NLP can be used for extracting knowledge from published medical papers. It acknowledges the fact those tools capable of identifying reliable information in the medical domain stand as building blocks for a healthcare system that is up-to-date with the latest discoveries. Our research focus on the diseases and treatment information and the relation that exists between these two entities.

Keywords: Healthcare, machine learning, natural language processing, Disease Treatment Extraction, Medline

I.INTRODUCTION

This Work provides the foundation for development of technology framework that makes easy to find all the relevant information regarding treatment and diseases. The tool that is built with the techniques such as Natural Language Processing (NLP) and Machine Learning (ML) has capability to find all relevant short text information regarding diseases and treatments. This work presents various Machine Learning (ML) and information for classifying short texts and relation between diseases and treatments. According to ML technique the information are shown in short texts when identifying relations between two entities such as diseases and treatment. Thus there is improvement in solutions when using a pipeline of two tasks (Hierarchical way of approaching). It is better to identify and remove the sentence that does not contain information relevant to disease or treatments. The remaining sentences can be classified according to the interest. It will be very complex 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.

Relation Extraction is a long standing research topic in Natural Language Processing. Medical information is stored in textual format among the biological data stored in Medline. Manually extracting useful information from large volume of database is a tedious work. Moreover HTML page displaying biological information contains medical information and typically unrelated materials such as navigation menus, forms, user comments, advertisement, feedback etc. The proposed work of this project extracts the useful disease related information with increased precision by using weighted bag of word representation [1] with an accuracy of 79% to 82%. The pro-posed approach supports in clinical decision making by providing physician with best available evidence of medical information. The frequent use of electronic health records and information increase the need for text mining in order to improve the quality of result for the user query. This can result in two area of real time application [7] such as Text search engine targeted with scientific document and Text Search engine targeted with technical document. In this project we choose text mining targeted with scientific document related to Medical treatment. Medline is chosen in this project to get biomedical information because it provides answers related to patient treatment and it's the database which is most widely used by the clinicians and research scholars in medical field. More importantly it is frequently updated and the contents are proved to be accurate compared to other medical websites providing information related to human disease, health, medicines, treatment etc. With the growing number of medical thesis, research papers, research articles, researchers are faced with the difficulty of reading a lot of research papers to gain knowledge in their field of interest. Search engines like Pub Med [8] reduces this constraint by retrieving the relevant document related to the user query.

Though the relevant document is retrieved, the web page displaying it may contain many non informative contents like advertisement, scroll bars, menus, citations, quick links, announcements, special credits, related searches, similar posts searched etc. This may be quite frustrating to the user when the user is in need of the information alone. In this project all the unrelated contents like advertisement etc mentioned in the above paragraph are removed and text mining is performed on the extracted document from which information or sentences related to user specified disease is extracted. From the extracted file symptoms, causes, treatment of the particular disease is filtered and displayed to the user. Thus the user gets the required information alone which saves his time and improves the quality of the result. This text mined document can be used in medical health care domain where a doctor can analyze various kinds of treatment that can be given to patient with particular medical disorder. The doctor can update the knowledge related to particular disease or its treatment methodology or the details of medicine that are in research for a particular disease. The doctor can gain idea about particular medicine that is effective for some patient but causes side effect to patient with some additional medical disorder. The patient can also use this extracted document to get clear understanding about a particular disease its symptoms, side effects, its medicines, its treatment methodologies.

Understanding the effect of a given intervention on the patient's health outcome is one of the key elements in providing optimal patient care. In the proposed approach a combination of structural natural language processing with machine learning method address the general and domain specific challenges of information extraction. Medical subheadings and subject heading may be used to infer relationship among medical concepts. The classification algorithm used in the proposed work exhibits effectiveness, efficiency, online learning ability.

II.LITERATURE SURVEY

In Jason D. M. Rennie, Lawrence Shih, Jaime Teevan, David R. Karger, "**Tackling the POOR Assumption of Naïve Bayes Text Classifier"** this paper described text classification by using naïve bayes text classifier but this text classifier does not give precision 100% for output. Sometimes prediction of classifier may be incorrect. [1]

In T.Mouratis, S.Kotsiantis, "**Increasing The Accuracy Of Discriminative Of Multinomial Bayesian Classifier In Text Classification",** there were introduced use of classifier that increased precision of output but the main problem in that work was at the time of classification it doesn't identify the verbs, nouns, adjectives, phrase properly so many time it may provided wrong data.[2]

In B.Rosario and M.A. Hearst**, "Semantic Relation in Bioscience Text"** In this paper author used Hidden Markov models for entity recognition. This includes mapping medical information into structural representation. It converts natural language text into structural format. Also they use machine learning for information extraction.

Text classification is used for the extraction of biomedical abstract. Semantic lexicons of words labeled with semantic classes so associations can be drawn between words which helps in extracting the necessary sentences related to the query. In this paper sentence co-occurrence and naive bays algorithm are used for extracting semantic relation like Gene-Protein from Medline abstract, the precision and recall of the result obtained are shown in the graph but due to use of only one naive bays algorithm it do not get good precision of output and it doesn't used bag of words to find adjective, verbs nouns phrase while doing classification. [3]

In M.Craven, **"Learning To Extract Relations from Medline"** In this research paper the individual sentences are considered as instances that are to be processed by the navie bayes classifier. Here each sentence is considered as positive training set. Relation extractions are made through relational learning. Extraction of words from Medline abstract has been done by using navie bayes and CNB algorithms. It also used bag of words during classification but it is not used natural language processing due to this performance of output degrades.[4]

In L. Hunter and K.B. Cohen**, "Biomedical Language Processing: What's Beyond Pubmed"** this system is used natural language processing for processing of biomedical words. In this work it takes the name of disease and give the solution which has been stored in local database of that disease by parsing user statement using natural language processing but it does not do diagnosis of disease.[5]

In Jeff Pasternack, Don Roth **"Extracting Article Text From Web With Maximum Subsequence Segmentation"** It involves to extract word according to occurrence of that word in article if no of word occur by no of time mentioned then extract that word from the web here author used bag of word to remove verbs and adjective from the article but it doesn't use Natural language processing while extracting.[6]

In Abdur Rehman, Haroon.A.Babri, Mehreen saeed, **"Feature Extraction Algorithm for Classification of Text Document"** It involves automatic extraction of semantic relation between medical related points. A dictionary of biomedical terms is used for sentence classification. The sentences are automatically parsed using semantic parser by using four classification algorithm such as NB, CNB, Decision tree, Adaptive, SVM etc while extracting word but it doesn't provided the information regarding diagnosis of disease.[7]

In Adrian Canedo-Rodriguez, Jung Hyoun Kim,etl., **"Efficient Text Extraction Algorithm Using Color Clustering For Language Translation In Mobile Phone"** Auther used AdaBoost classifier is outperformed by other classifier. SVM classifier is always functions well when the information matches with the training set. Probabilistic model are used to perform text classification task. Bag of word technique is simple in nature but in many time it is hard to outperform it. Pipelining task is essential to obtain increased quality of result because majority functions may overcome the underrepresented ones. By using pipelining there is a balance between relevant and irrelevant data and the classifier has better chance to distinguish relevant and irrelevant data but it don't used Gennia tagger tool which is special parser for biomedical words.[8]

In Oana Frunza, Diana Inkpen, and Thomas Tran, Member, IEEE "**A Machine Learning Approach for Identifying Disease-Treatment Relations in Short Texts**" In this work it involves two task in pipelined manner for identifying and extracting the relationship between the given MEDLINE abstract. Firs 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 algorithms namely decision based models, probabilistic models, Adaptive learning, linear classifier like support vector machine and a classifier that always predicts the majority class in training data are used. 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 but it does not give disease diagnosis as well information about particular disease by parsing statement.[9].

III.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

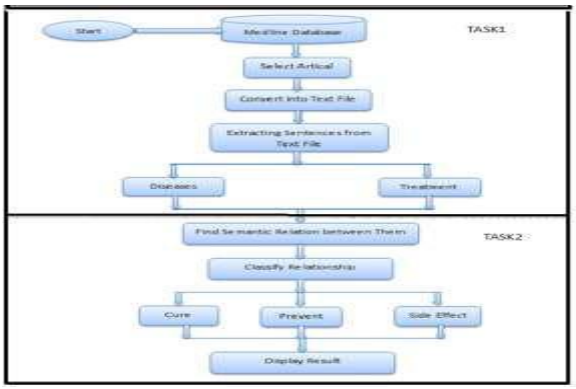

Fig 1: System Architecture

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].

IV.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.

V.CONCLUSION

This approach is very useful for everyone as it gives in-formation only of the area of interest. The interests are in line with the tendency of having a personalized medicine that has one in which each patient has its medical care tailored to its needs. This study is related to a particular field but the future scope of the paper lies in the fact that this can be extended to the information on the web. The proposed system used the top concept candidate for each identified phrase in an abstract as a feature. Experimental result shows that the technique used in

the proposed work minimizes the time and the work load of the doctors in analyzing information about certain disease and treatment in order to make decision about patient monitoring and treatment. This text mined document can be used in medical health care domain where a doctor can analyze various kinds of treatment that can be given to patient with particular med-ical disorder. The doctor can update the knowledge re-lated to particular disease or its treatment methodology or the details of medicine that are in research for a particular disease. The doctor can gain idea about particular medi-cine that are effective for some patient but causes side ef-fect to patient with some additional medical disorder. The patient can also use this extracted document to get clear understanding about a particular disease its symptoms, side effects, its medicines, its treatment methodologies. This paper also present healthcare diagnosis treatment & prevention of disease, illness, injury in human.

REFERENCES

[1] Jason D. M. Rennie, Lawrence Shih, Jaime Teevan, David R. Karger, "Tackling the POOR Assumption of Naïve ayes Text Classifier", Proceedings Of The Twentieth International Conference On Machine Learning (ICML-2003), Washington DC, 2003.

[2] T.Mouratis, S.Kotsiantis, "Increasing the Accuracy of Discriminative of Multinomial Bayesian Classifier in Text Classification", Classification", ICCIT'09 Proceedings Of The 2009 Fourth International Conference On Computer Science and Convergence Information Technology

[3] B.Rosario and M.A.Hearst, "Semantic Relation in Bioscience Text", Proc. 42nd Ann. Meeting on Assoc for Computational Linguistics, Vol.430, 2004.

[4] M.Craven, "Learning To Extract Relations From Medline", Proc. Assoc. For The Advancement Of Artificial Intelligence.

[5] L.Hunter and K.B.Cohen, "Biomedical Language Processing: What's Beyond Pubmed?" Molecular Cell, Vol. 21-5 Pp. 589-594, 2006.

[6] Jeff Pasternack, Don Roth "Extracting Article Text from Webb with Maximum Subsequence Segmentation", bb, WWW 2009 MADRID.

[7] Abdur Rehman, Haroon.A.Babri, Mehreen saeed," Feature Extraction Algorithm For Classification Of Text Document", ICCIT 2012 .

[8] Adrian Canedo-Rodriguez, Jung Hyoun Kim,etl.," Efficient Text Extraction Algorithm Using Color Clustering For Language Translation In Mobile Phone" , May 2012.

[9] In Oana Frunza, Diana Inkpen, and Thomas Tran, Member, IEEE "A Machine Learning Approach for Identifying Disease-Treatment Relations in Short Texts"May2011

[10] 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,

[11] 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.

[12] National Center for Biotechnology Information. Entrez ProgrammingUtilities Help, 2010.

[13] B.J. Stapley and G. Benoit, ―Bibliometrics: Information Retrieval Visualization from Co-Occurrences of Gene Names in MEDLINE Abstracts,‖ Proc. Pacific Symp. Biocomputing, vol. 5, pp. 526-537, 2000.