Volume 3, Issue 3, March 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Efficient Extraction of Medical Relations using Machine Learning Approach

Janani.R.M.S
M.E Computer Science And Engineering,
Kalasalingam Institute Of Technology,
Virudhunagar, India.

Ramesh.V
Assistant Professor,Dept CSE,
Kalasalingam Institute Of Technology,
Virudhunaga,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 we use Multinominal Navie Bayes algorithm and several other techniques to extract semantic relation between disease and their associated treatment with increased rate of Precision, Recall and F-measure. 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. The proposed technique can be integrated with any medical management system to make better medical decisions and in patient management system by automatically mining the biomedical information from digital repositories like Medline.

Keywords— Disease Treatment Extraction, Increased Precision, Machine Learning, Medical Care Domain, Stemming Algorithm and Sequence Pattern.

I. INTRODUCTION

Relation Extraction is a long standing research topic in Natural Language Processing. Medical information are 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 a accuracy of 79% to 82%. The proposed 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 inorder 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 analyse 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 are 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

The most relevent work is the work done by Rosario and Hearst ***[4]*** where Hidden Markov models are used for entity recognition. This includes mapping biomedical information into structural representation. It involves converting natural language text into structural format. Their work uses machine learning for information extraction. The extraction of medical abstract is obtained through text classification. 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 research paper the author used sentence co-occurrence and navie bayes algorithm to extract semantic relation like Gene-Protein from medline abstract, the precision and recall of the result obtained are shown in the graph as their experimental results.

In their work the individual sentences are considered as instances that are to be processed by the navie bayes classifier. Here each instance is considered as positive training set. Alternative relation extraction are made through relational learning. Relational learning ***[3]*** involves parsing sentences and from the parsed sentences, parse tree is constructed. From the parsed tree grouping of the relevant sentence is made. The extracted results are in the structured form. The task of relation extraction was previously tackled in medical literature for gene-disorder association **[5]**. It involves automatic extraction of relation between medical concepts. A dictionary of medical terms are used for sentence classification. The sentences are automatically parsed using semantic parser. After applying semantic extraction a set of extraction, alteration, validation rules are applied to distinguish the actual semantic relation to be extracted. Protein interaction information is hidden in most of the medical abstracts. Interaction extraction approach is used to identify protein-protein relation. First dependency relation of the sentence is made, this helps us to identify the protein sentences rather than using classical sentence relation identification. Semi-supervised machine learning methods are used on the dependency features selected from the first step. By using semi-supervised learning method the cost of labelling is reduced significantly.

Kernel method can be used when the situation is computationally infeasible because of infinite free dimension. Kernel based approach gives new methods for relation extraction regarding syntactic information in which parse tree is constructed and are then compared by the kernel function for correctness. It results in bio-entity recognition which increases the interest for extracting biomedical relation. Kernel uses subsequent pattern to confirm relation between two entities.

Oana Frunza.et.al ***[6]*** in her work performed 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 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 all the 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.

AdaBoost classifier is outperformed by other classifier ***[9].*** SVM classifier always functions well when the information matches with the training set. Probabilistic model always performs well on text classification task. Bag of word technique is simple in nature and in majority of the cases it is hard to outperform it. Pipelining of task is essential to obtain increased quality of result because majority class 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 non-relevant data.

III. PROPOSED SYSTEM

Before we 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 filewith .html extension as .txt extension.

This invoves 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 algorithm, association rules and representation technique to achieve a processed text file which contains only the information related to Symptoms, Causes, Treatment about the disease in the user spcified .html document. The application is designed using java where java swing is used to create a GUI in which the user uploads the .html

document. Two buttons are designed with the label of extract and remove. When extract button is pressed the .html file content is converted into a text fileby eliminating the html tags and extracting only the textual content from the html file.

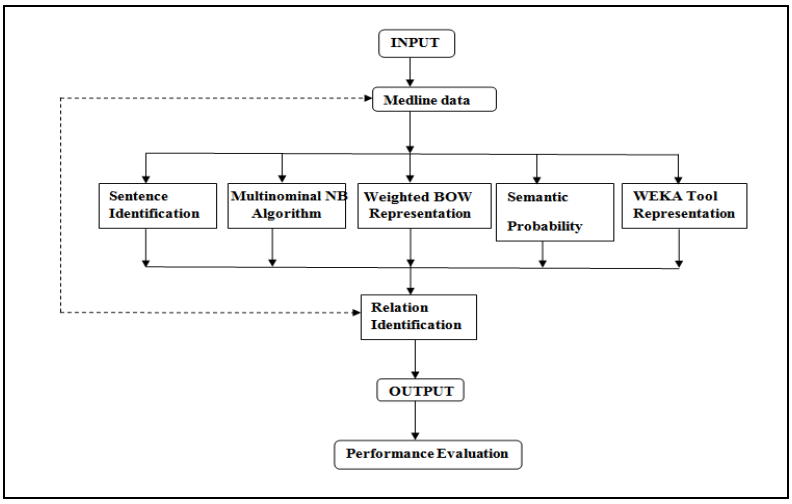

Figure 1: System Archietecture Of The Proposed System

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 endind with ing, ed etc. These words can be removed to improve the quality of the result. Thus stopwords are removed from the extracted text file.

Now the stop word removed text file is subjected to the combination of certain words inorder to avoid repetation such as excessed, excessing if both these words apper 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 repeatation of words for user to get clear understanding of sentences or information. By removing such suffixes and combining these kind 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. Inorder to extract this semantic relations a classification algorithm namely Multinominal Naïve Bayes classification algorithm is used in association with Aprior association rule mining. The reason for choosing Multinominal NB and the drawback of Naïve Bayes algorithm are discussed below.

Multinominal Naïve Bayes is the specialized version of Naïve Bayes specially used for text documents. Multinominal NB models the word count and performs the classification within it. Aprior association mining is used to find co-occurance of features in the form of association rules. Textual representation technique for labelling the training data and for identifying the sentences related to the label Symptoms, Causes and Treatment are achieved by using Weighted Bag-Of-Word representation technique along with word sequence pattern is used.

Word sequence pattern approach is used to analyze data and identify the pattern, such patterns can be used to make prediction which is an effective step in decission making. It can be applied to identify pattern in health care domain to find pattern observed in the symptoms of particular disease. Now the resulted file containing information related to Symptoms, Causes, 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 obtained result is assumed to have quality if these values are within the range of 0.0 to 1.0.

The formulas for calculating these quality measures are Precission= (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. The proposed system is validated using html page containing information related to lung cancer and the screen shots are displayed in validating the result section. The modules designed to achieve the proposed idea are as follows :

A. Html To Text Conversion

The saved .html document is converted into a text file and is stored with .txt extension. The stored text file contains Disease-Treatment relation along with the details of the code involved in designing the page, forms, suggestion box, navigation menus, advertisement, feedback, etc. In this text file, the classification algorithm and representation techniques and other analysis algorithms are used to obtain disease-Treatment relation with high precision pattern to minimize the number of false positive relation extraction.

B. 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. BOW represents a document as a histogram of word occurrences. Such representation is unable to maintain any sequential information. In the proposed work Weighted Bag-Of-Word representation is used which overcomes the above mentioned problem of BOW.

1) **Weighted Bag-Of-Word Representation:**

Weighted BOW uses local smoothing to embed documents as smooth curves in the multinomial simplex thereby preserving valuable sequential information. Weighted BOW is able to robustly capture medium and long range sequential trends in the document.

C. Sentence Identification And Relationship Extraction

The sentence identification and relation extraction task involves identifying Disease sentences and its treatment relationship. Naïve Bayes classifiers are fast and easy to implement but affects the quality of the result. Two reasons for Naïve Bayes poor performance[2] are

i) Naïve Bayes selects poor weight for decision boundary. To balance the amount of training example used,"compliment class" formulation of Naïve Bayes can be used.

ii) Features are assumed to be independent. Weight of class with stronger word dependencies is greater than class with weak word dependencies. To prevent this domination, classification weight can be normalized.

1)Multinominal Naïve Bayes Classification Algorithm:

To resolve the above problem and to result in efficient sentence identification Multinominal Naïve Bayes classification algorithm is used in the proposed system. Multinominal Naïve Bayes classification (MNB) **[3]**algorithm adopts parameter learning method. Disease sentence are identified using MNB algorithm in which performance of the classifier is improved by adopting certain features of Compliment Naïve Bayes Classifiers.

***2)Aprior Association Rule Mining*:**

Aprior association rule mining technique extracts the useful relations **[3].** It discovers the relations between variables in large database. It must satisfy the user specified minimum support count and minimum confidence level. This information obtained can be used for making decision about patient's treatment and the new kind of medicine or treatments under research.

D. Output Performance Evaluation

The performance of the proposed model is tested with the html page from MEDLINE containing information about lung cancer. The end result was a text file containing only the information about the lung cancer disease mainly on symptoms, causes and treatment with increased rate of precision compared to the .html file which was given as input.

The .html file and the resulted extracted text file is calculated for precision, recall and F-measure. A graphical representation is used to represent the comparison results based on word count.

IV. VALIDATING THE PROPOSED SYSTEM

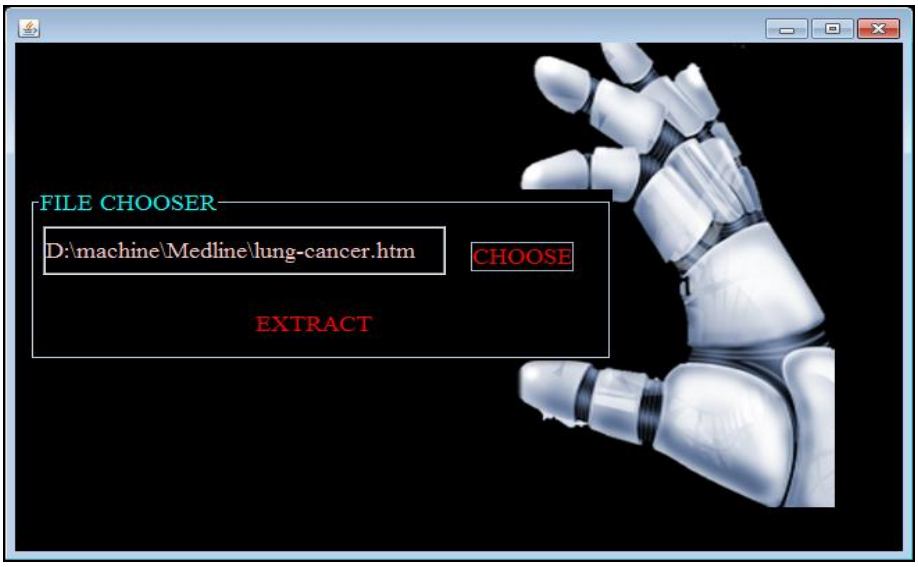

Figure 2: Uploading .html document

Figure 2 represents uploading of a .html file containing information about lung cancer from MEDLINE.

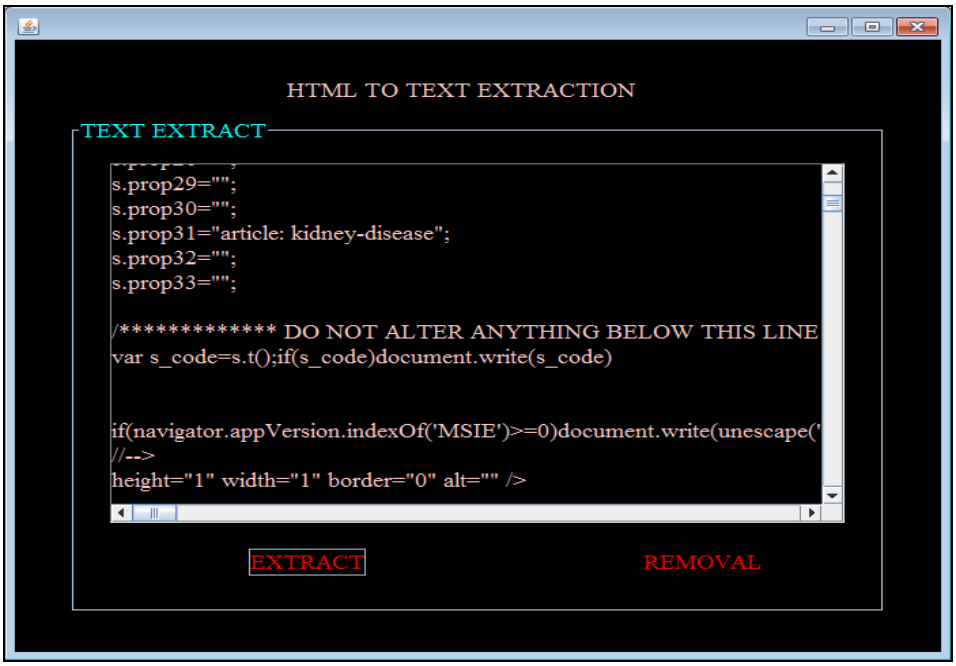

Figure 3: HTML to Text Conversion

Figure 3 shows the conversion of the uploaded .html document to the text document by removing all the html tags and by clicking the REMOVAL button all the stopwords in the text file is removed to improve the quality of the file by removing the redundant words.

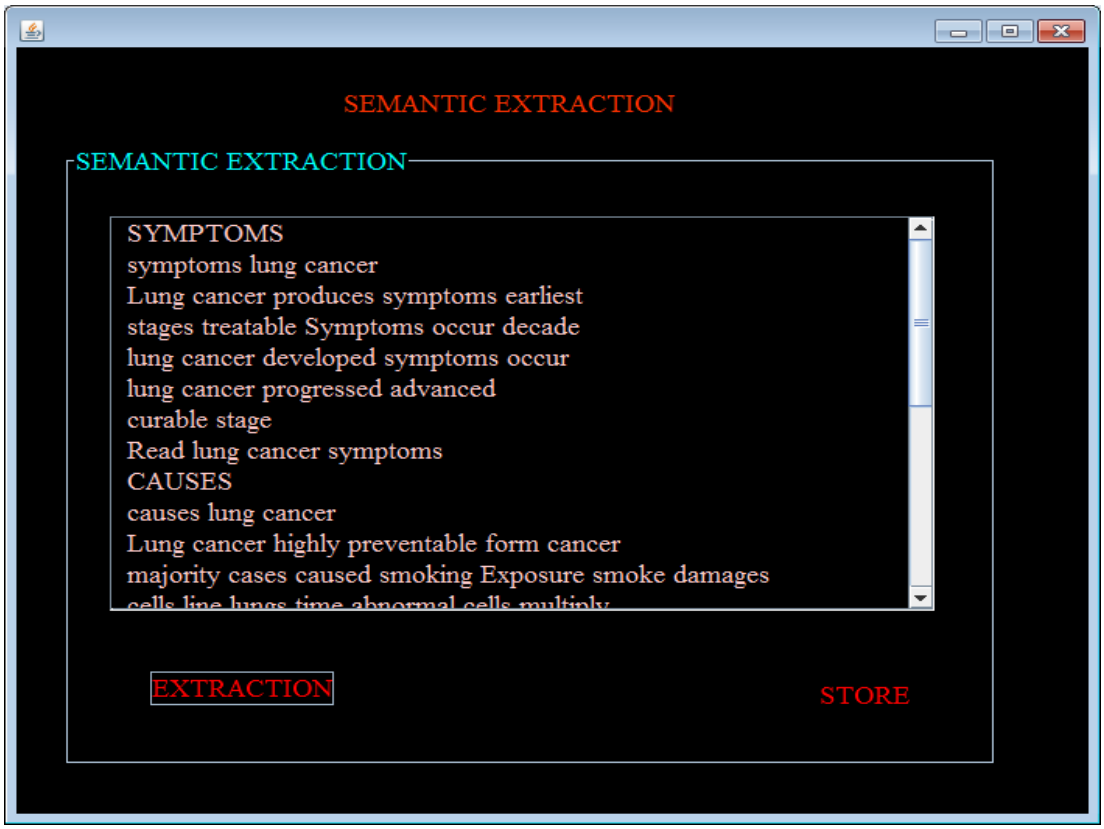

Figure 4: Disease-Treatment Relation Extracted Using MNB

Figure 4 shows the extraction of Disease-Treatment relation using Multinominal Naïve bayes algorithm and Aprior association rule mining algorithm and representation techniques like Weighted Bag-Of-Word, Word sequence pattern approach. After semantic extraction is performed, the resulted text file contains information about Symptoms, Disease, Treatment related to lung cancer disease from the uploaded html file. Extracted file is stored as a text file with the name of the disease namely lungcancer.txt for user to directly read it.

Figure 5 shows the quality of the extracted file. The quality of the extracted file is calculated by counting the number of word count from the extracted information under heading of Symptoms, Causes, Treatment.

Then the calculated word count is used in the formula of Precision, Recall and F-measure. It is shown that the resultant values are between 0.0 to 1.0 which assures the quality of the extracted file containing only the necessary information compared to the input html file with lots of irrelevant informations.

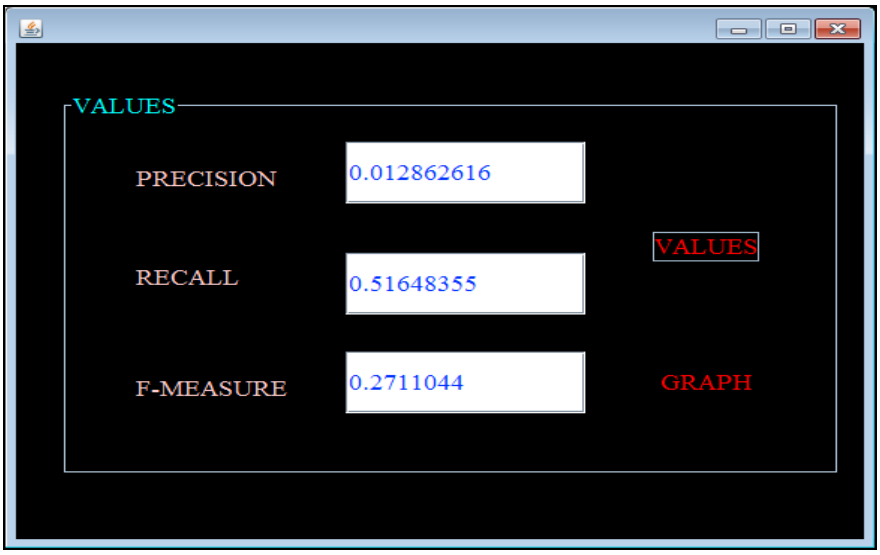

Figure 5: Calculated Metrics Of Extracted File

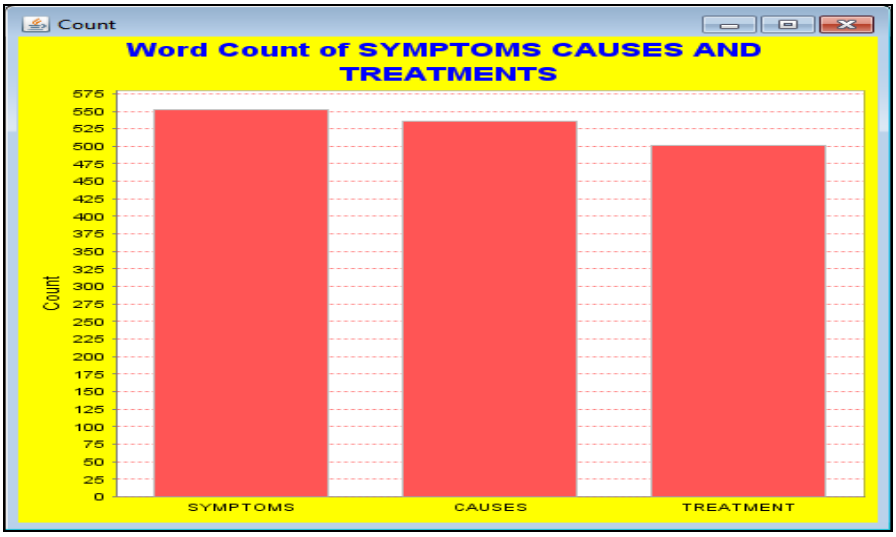

Figure 6:Bar chart Representation Of Extracted Output File

Figure 6 shows the bar chart representation of the extracted text file saved as lungcancer.txt containing information only about Symptoms, Causes, Treatment of the uploaded lung cancer html file. It depicts the number of word count under each category in the extracted lung cancer text file.

V. CONCLUSION

The proposed system removes the unwanted contents from the HTML page from MEDLINE and result on a text document containing only the particular disease and its relevant Symptoms, Cause and Treatment. 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 analyse 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 are 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. Future enhancement is to extend this work to extract Disease-treatment relation from certain medical database or search engine.

REFERENCES

[1] Jason D. M. Rennie, Lawrence Shih, Jaime Teevan, David R. Karger ,***"Tackling The POOR Assumption Of Naïve Bayes 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 Multinominal Bayesian Classifier In Text 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] Oana Frunza.et.al, "***A Machine Learning Approach For Identifying Disease-Treatment Relations In Short Texts",*** May 2011

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

[7] Jeff Pasternack, Don Roth ***"Extracting Article Text From Webb With Maximum Subsequence Segmentation",*** WWW 2009 MADRID.

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

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

Janani.R.M.S, is presently doing her final year Master Of Engineering in Kalasalingam Institute Of Technology, Virudhunager. Her research interest includes Data Mining, Knowledge Engineering .

V.Ramesh, currently working as an Assistant Professor in the department of Computer science and Engineering in Kalasalingam Institute of Technology, Virudhunager. He has presented papers in International and National conferences . His area of interest are Data Mining and Cloud Computing.

