RESEARCH ARTICLE OPEN ACCESS

Smart health Analytics using Natural Language Processing and Machine Learning

Jyostnarani Tripathy[1], Santosh Ku. Satapathy[2], Dr.Sujit Panda[3]

[1,3]Associate Professor, Department of Computer Science Engineering, Gandhi Institute For Technology (GIFT), Bhubaneswar

[2] Assistant Professor, Department of Computer Science Engineering, Gandhi Engineering College, Bhubaneswar

ABSTRACT: This Paper presents efficient machine learning algorithms and techniques used in extracting disease and treatment related sentences from short text published in medical papers. The main objective of this work is to show what Natural Language Processing (NLP) and machine learning techniques used for representation of information and what classification algorithms are suitable for identifying & classifying relevant medical information in short text. This paper also present healthcare diagnosis treatment & prevention of disease, illness, injury in human. The domain is automatically learn some task of healthcare information, medical management, Patient health information etc. The proposed technique can be integrated with any medical management system to make better medical decision and in patient management system can automatically mining biomedical information from digita lrepositories.

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 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 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 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. LITERATURESURVEY

[1] In Jason D. M. Rennie, Lawrence Shih, Jaime Teevan, David R. Karger ,**"Tackling The POOR Assumption Of Naïve Bayes Text Classifier"** there were mentioned classification of text by using naïve bayes text classifier but use of navie bayes text classifier does not give precision 100% for output. Sometimes prediction of classifier may not becorrect.

[2] In T.Mouratis, S.Kotsiantis, "**Increasing The Accuracy Of Discriminative Of Multinominal Bayesian Classifier In Text Classification",** paper author introduced use of classifier that increased precision of output but problem in that work was at the time of classification it doesn't identify the verbs,nouns,adjectives properly so some time it nay give wrongvalue.

[3] In B.Rosario And M.A. Hearst**,"Semantic Relation In Bioscience Text"** 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 but due to use of only one navie bayes algorithm it do not get good precision of output, it doesn't used bag of words to find adjective, verbs while doing classification. [4]In M.Craven, **"Learning To Extract Relations From Medline"** 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. Extraction of words from medline abstract has been done by using navie bayes,CNB algoritham and it also used bag of words during classification but not used natural language processing due to this performance of outputdegrades.

[5] In Oana Frunza.et.al, "**A Machine Learning Approach For Identifying Disease-Treatment Relations In Short Texts"** It involves automatic extraction of relation between medical concepts. A dictionary of medical terms is 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 but problem is that due to used of only one algorithm of machine learning navie bayes may not get good precision ofoutput.

[6] In L. Hunter And K.B. Cohen**, "Biomedical Language Processing: What's Beyond Pubmed"** it involves Used of 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 database of that disease by parsing user statement using natural language

processing but it does not do diagnosis ofdisease.

[7] In Jeff Pasternack, Don Roth **"Extracting Article Text From Webb With Maximum Subsequence Segmentation"** in involves to extract word according to occurrence of that word in article if no of word occur by no of time mentioned then I 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 whileextracting.

[8] In Abdur Rehman, Haroon.A.Babri, Mehreen saeed,**"Feature Extraction Algorithm For Classification Of Text Document"** It involves automatic extraction of relation between medical concepts. A dictionary of medical terms is used for sentence classification. The sentences are automatically parsed using semantic parser.it use 4 classification algorithm NB,CNB,Decissiontree,Adaptive ,SVM etc wile extracting word but it doesn't give the information regarding diagnosis of disease.

[9] In Adrian Canedo-Rodriguez, Jung Hyoun Kim,etl**.,"Efficient Text Extraction Aalgorithm Using Color Clustering For Language Translation In Mobile Phone"** AdaBoost classifier is outperformed by other classifier**.** 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 but it don't used Gennia tagger which is special parser for medical words.

[10] In Oana Frunza, Diana Inkpen, and Thomas Tran, Member, IEEE "**A Machine Learning Approach for Identifying Disease-Treatment Relations in Short Texts**" 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 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 but it does not give disease diagnosis as well information about particular disease by parsingstatement.

III. PROBLEMSTATEMENT

According to Literature survey of above 10 papers some of papers have features but also some problems. So i have combined all the features like extracting word from Medline from **Jeff Pasternack [7], Don Roth, Oana Frunza[5],**Used of six machine learning algorithm in order to improve precision of output NB,CNB,Adaptive Boost,SVM,Decission Tree,BaseLine **Oana Frunza, Diana Inkpen, and Thomas Tran, Member [10],**Used of diagnosis of disease **M.Craven [4],**It also gives information about prevention,cure,treatment about any diseases from database by parsing user query through natural Language processing **L. Hunter And K.B. Cohen** [6],Used of Genia tagger which is special parser for medical words **L. Hunter And K.B. Cohen** [6] in my project.

IV. PROPOSEDSYSTEM

The two tasks used in this paper are the basis for the development of information technology framework. This framework helps to identify the medical related information from abstracts. The first task deals with extraction all information regarding diseases and treatments while the task deals with extraction of related information existing between disease and treatments. The framework developed with these tasks are used by healthcare providers, people who needs to take care of their health related problems and companies that build systematic views. The future product can be provided with browser plug-in and desktop application so that it helps the user to get all information related to diseases and treatments and also the relation between those entities. It is also be useful to know more about latest discoveries related to medicine. The product can be developed and sold by companies that do research in medical care domain, Natural Language Processing (NLP), and Machine Learning (ML), and companies that develop tools like Microsoft Health Vault and Google Health. This product is valuable in e-commerce fields by showing the statistics that the information provided here are accurate and also provide all the recent discoveries related to health care. To make a product more popular it should be trust worthy so that people can buy it. It is the key factor foe any company to make product successful. When coming to health care products it should be more trust worthy since it is dealing with health related problems. Companies that wish to sell health care framework need to develop tools that automatically extract the wealth of research. For example the information provided for diseases or treatments needs to be based on recent discoveries on health care field so that people can trust. The product quality also should be taken care so that it provides dynamic content for users. The first task deals with the identification of sentences from the Medline abstracts that provide the information about the diseases and treatments. In other words it also seems like scanning the sentences from Medline abstracts that contain relevant information which the user wants. Natural Language Processing (NLP), and Machine Learning (ML) are used to extract accurate information or it can also say that it perfectly removes the unwanted information which are not related to disease or treatment. Natural Language Processing (NLP) and Machine Learning (ML) itself involve in extracting informative sentences. It is difficult task to identify the informative sentences in fields such as summarization and information extraction. The work and contribution value with this task is helpful in results and in settings for this task in healthcarefield.

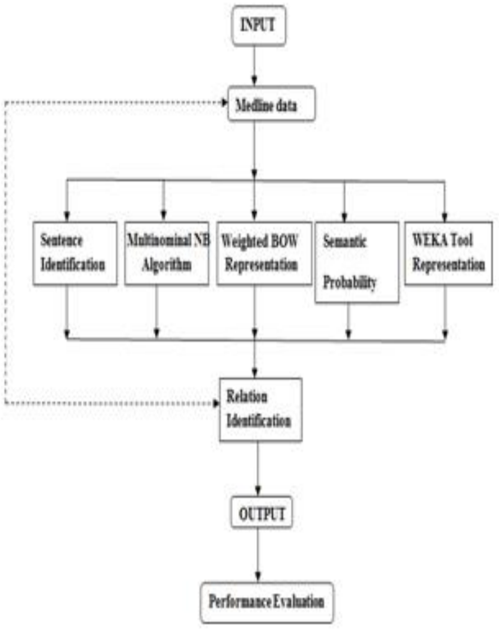


Figure 1: System Architecture Of The Proposed System

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. In order to extract this semantic relations a classification algorithm namely Multinomial Naïve Bayes classification algorithm is used in association with Aprior association rule mining. The reason for choosing Multinomial NB and the drawback of Naïve Bayes algorithm are discussed below. Multinomial Naïve Bayes is the specialized version of Naïve Bayes specially used for text documents. Multinomial NB models the word count and performs the classification within it. A prior association mining is used to find co-occurrences 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 approachis used to analyze data and identify the pattern, such patterns can be used to make prediction which is an effective step in decision 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, 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 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 Precision= (relevant retrieved)document/ Retrieved document. Recall =(relevant retrieved) document/ Relevant document. F-measure= mean of Precision andRecall.

4.1 Task and DataSets

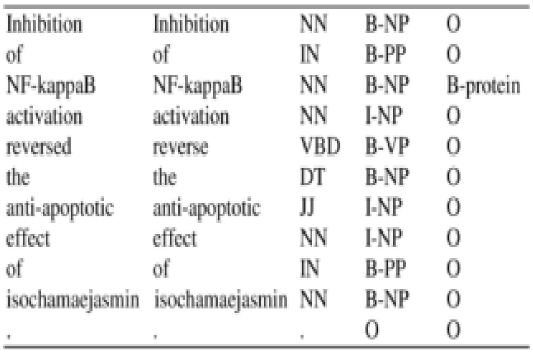
The two tasks that are undertaken in this paper provide the basis for the design of an information technology framework that is capable to identify and disseminate healthcare information. The first task identifies and extracts informative sentences on diseases and treatments topics, while the second one performs a finer grained classification of these sentences according to the semantic relations that exists between diseases and treatments.

4.2 Bag OfWords

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. Two most common feature value representations for BOW representation are: binary feature values—the value of a feature can be either 0 or 1, where 1 represents the fact that the feature is present in the instance and 0 otherwise; or frequency feature values—the value of the feature is the number of times it appears in an instance, or 0 if it did not appear. Because we deal with short texts with an average of 20 words per sentence, the difference between a binary value representation and a frequency value representation is not large. In our case, we chose a frequency value representation. This has the advantage that if a feature appears more than once in a sentence, this means that it is important and the frequency value representation will capturethis.

4.3 GeniaTagger

Type of representation is based on syntactic information: noun-phrases, verb-phrases, and biomedical concepts identified in the sentences. In order to extract this type of information, we used the Genia11tagger tool. The tagger analyzes English sentences and outputs the base forms, part-of-speech tags, chunk tags, and named entity tags. The tagger is specifically tuned for biomedical text such as Medline abstracts.



Inhibition	Inhibition	NN	B-NP	O
of	of	IN	B-PP	O
NF-kappaB	NF-kappaB	NN	B-NP	B-protein
activation	activation	NN	I-NP	O
reversed	reverse	VBD	B-VP	O
the	the	DT	B-NP	O
anti-apoptotic	anti-apoptotic	JJ	I-NP	O
effect	effect	NN	I-NP	O
of	of	IN	B-PP	O
isochamaejasmin	isochamaejasmin	NN	B-NP	O
.	.	.	O	O

Fig. 2. Example of Genia tagger output including for each word: its base form, its part-of-speech, beginning (B), inside (I), outside (O) tags for the word, and the final tag for the phrase.

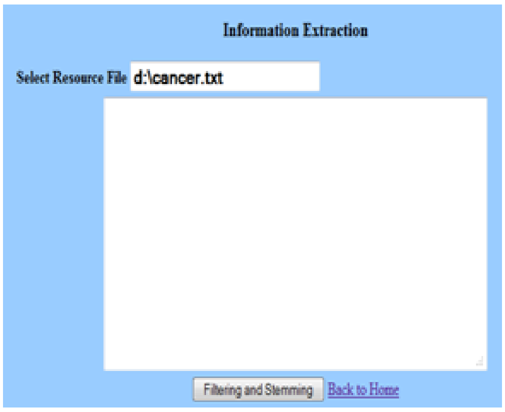
Fig.2presentsan example of the output of the Genia tagger for the sentence: “Inhibition of NF-kappaB activation reversed the anti-apoptotic effect of is ochamaejasmin.” The noun and verb-phrases identified by the tagger are features used for the second representation technique. We ran the Genia tagger on the entire dataset.

4.4 DiagnosisAlgorithm

I/P
D: a Training Set. N:Number of instance O/P. F:FilteredDataset.
O:OutlierDataset.

1. Empty F &O.
2. Train(T).
3. Assigni=1
4. If Dw € Tthen
5. Inser Dw to Felse
6. Insert Dw to Oend
7. Increase I by 1,then go to step4
8. Do it until i=N then go to step8
9. ReturnF,O

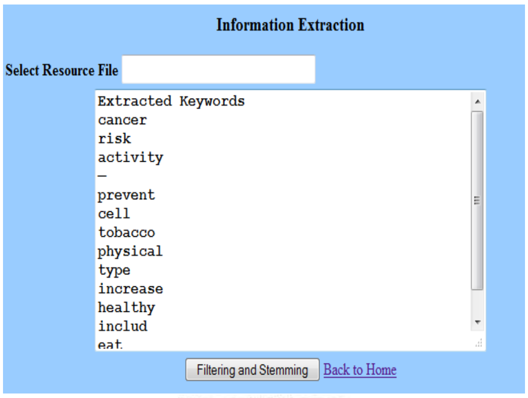
V. RESULTS AND EVALUATION



Information Extraction

Select Resource File d:\cancer.txt

Fig3.Select abstract file to extract information in short Text.



```
Extracted Keywords
cancer
risk
activity
_
prevent
cell
tobacco
physical
type
increase
healthy
includ
eat
```

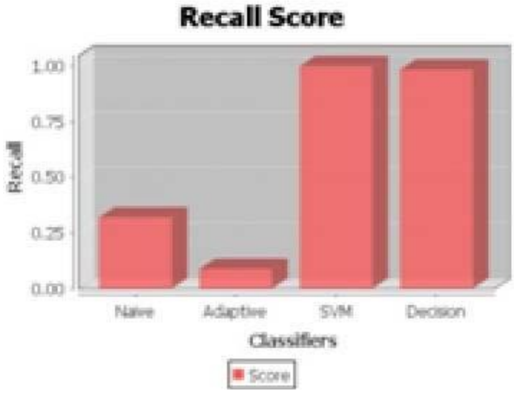


Fig 4.Extract information in short text from file selected in fig 1.

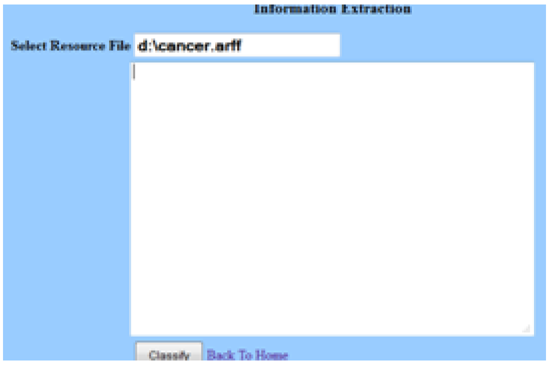
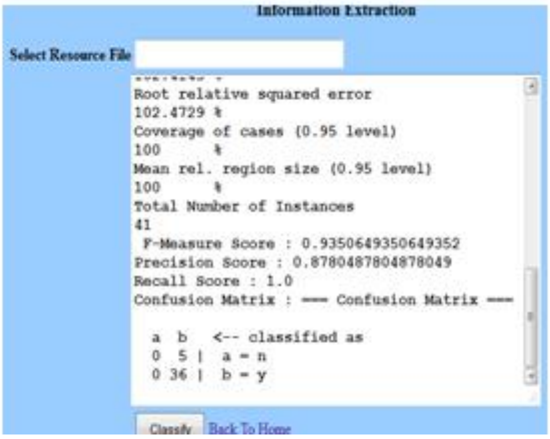


Fig 5.Select abstract file in arff format to classifier



```
Root relative squared error
102.4729 %
Coverage of cases (0.95 level)
100      %
Mean rel. region size (0.95 level)
100      %
Total Number of Instances
41
 F-Measure Score : 0.9350649350649352
Precision Score : 0.8780487804878049
Recall Score : 1.0
Confusion Matrix : === Confusion Matrix ===

  a  b   <-- classified as
  0  5 |  a = n
  0 36 |  b = y
```

Fig 6.Apply classification algorithm Naive Bayes and find F- measure,Precision,Recall.

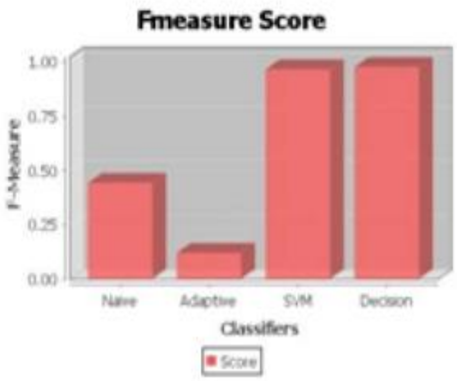


Fig 7. Fmeasure barchart for cancer abstract using NB,SVM,DicisionTree,AdaptiveLearning Classifier.

Fig 8. Recall barchart for cancer abstract using NB,SVM,DicisionTree,AdaptiveLearningClassifier.

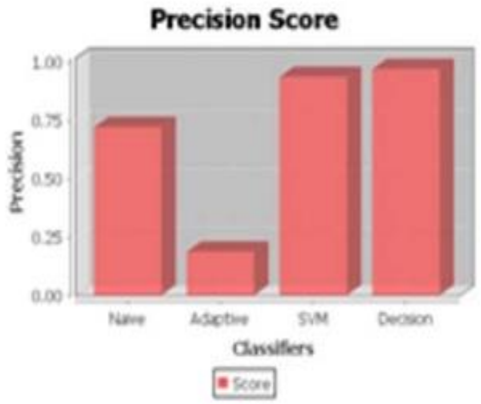


Fig 9. Precision barchart for cancer abstract using NB,SVM,DicisionTree,AdaptiveLearningClassifier.

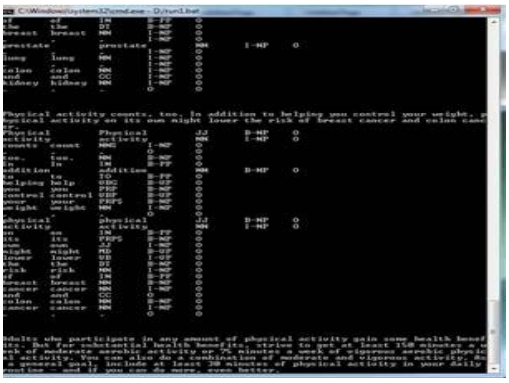
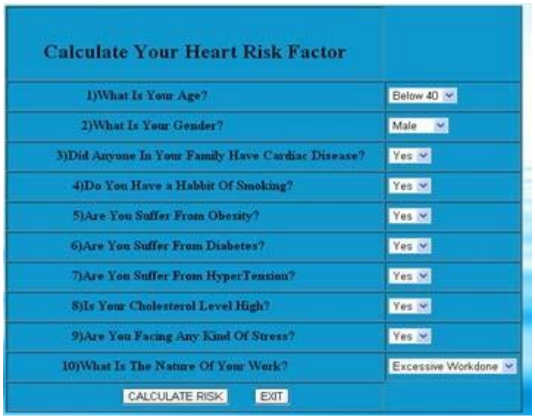


Fig 10. output of the Genia tagger for above cancer abstract that gives chunk of words.



Calculate Your Heart Risk Factor	
1)What Is Your Age?	Below 40
2)What Is Your Gender?	Male
3)Did Anyone In Your Family Have Cardiac Disease?	Yes
4)Do You Have a Habbit Of Smoking?	Yes
5)Are You Suffer From Obesity?	Yes
6)Are You Suffer From Diabetes?	Yes
7)Are You Suffer From HyperTension?	Yes
8)Is Your Cholesterol Level High?	Yes
9)Are You Facing Any Kind Of Stress?	Yes
10)What Is The Nature Of Your Work?	Excessive Workdone

Fig 11.Dignosis of Disease and find Risk factor by making Training Data Set.

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.

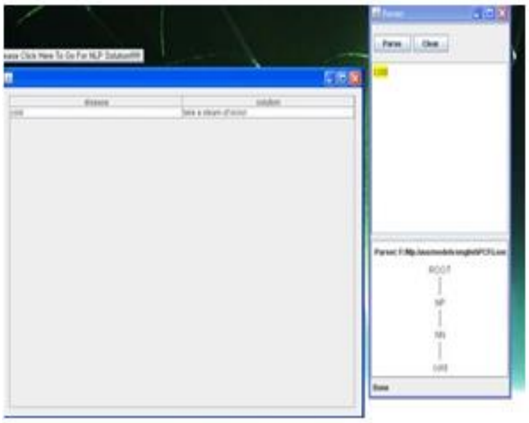


Fig 12.Find Disease Solution,Side effect,prevention by parsing user statement using NLP(Natural Language Processing.

VI. CONCLUSION

The proposed system removes the unwanted contents from the abstract 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 alsouse this extracted document

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 ArtificialIntelligence.

[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",** ICCIT2012.

[9] Adrian Canedo-Rodriguez, Jung Hyoun Kim,etl.**,"Efficient Text Extraction Aalgorithm Using Color Clustering For Language Translation In Mobile Phone"** , May2012.

[10] Oana Frunza, Diana Inkpen, and Thomas Tran, Member, IEEE "**A Machine Learning Approach for Identifying Disease-Treatment Relations in Short Texts**" IEEE transactions on knowledge and data engineering, vol. 23, no. 6, june2011