IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 3, Ver. VII (May-Jun. 2014), PP 05-12
www.iosrjournals.org

Best Treatment Identification for Disease Using Machine Learning Approach in Relation to Short Text

[1] Bharti E. Nerkar, [2] Sanjay S. Gharde

[1] *Research Scholar Dept. Computer Science, S.S.B.T.'s College of Engg.& Technology, Bambhori, Jalgaon, India*
[2]*Assistant Professor Dept. Computer Science, S.S.B.T's College of Engg. &Technology, Bambhori, Jalgaon, India*

Abstract: *The goal of Machine Learning is to construct a computer system that can adapt and learn from their experience. Machine Learning approach helps to integrate the computer based system into the healthcare field in order to obtain best and accurate results for the system. Here the system deals with automatic identification of informative sentences from medical published by medical journals. Our main aim is to integrate machine learning in medical field and build an application that is capable of automatically identifying and disseminating disease and treatment related information, further it also identifies semantic relations that exists between diseases and treatments. In the proposed work user will search for the disease summary (disease and treatment related information) by giving symptoms as a query in the search engine. Initially when a pdf is downloaded and saved in the system it first performs per processing on the data in the document and the extracted relevant data is stored in the database. The symptoms entered by the user are further classified using SVM classifier to make the further process easier to find the semantic keyword which helps to identify the disease easily and quickly. Then the semantic keyword found is matched with the stored medical input database to identify the exact disease related to that keyword present. Once the disease related to the symptom is identified, it is sent to medical database to extract the articles pertaining to that disease. The preprocessing process involves tokenization, removal of stop words and stemming. Followed by that, relevant information is extracted using the keyword searching algorithm. The combination of BOW, NLP and biomedical concepts are put together toe identifying semantic relations that exist between diseases and treatments in biomedical sentences. Till now the best result obtain is 98.51% F-measure by OanaFrunza, for the extraction of cure and prevents relations. In our implementation of our proposed system we have used SVM classifier which gives us an improved result. The problem statement of the existing system was, it didn't identify the best disease treatment. So the proposed solution used data mining concepts using voting algorithm to resolve the problem and find the best treatment for disease out of the treatment identified by the system.*

Keywords: *Datasets, Medline, HealthCare, Diseases and Treatment relation, Keyword Searching Algorithm, Rule Based classifier, Ranking Algorithm.*

I. Introduction

Research in the fields of life-science and biomedical domain has been the focus of the Natural Language Processing (NLP) and Machine Learning (ML) community for some time now. This trend goes very much in line with the direction the medical healthcare system is moving to the electronic world. The research focus of scientists that work in the field of computational linguistics and life science domains also followed the trends of the medicine that is practiced today, an Evidence Based Medicine (EBM). This new way of medical practice is not only based on the experience a healthcare provider acquires as time passes by, but on the latest Discoveries as well.

Machine Learning has emerged as an important technology almost in all domains of scientific research and medical fields. Machine Learning approach helps to integrate the computer system into the healthcare field in order to obtain best and accurate results. This paper deals with automatic identification of informative sentences from medical abstracts published by MEDLINE and gives out the medical measures that were insisted in that abstract such as best treatment for the disease found from the symptoms' given as input.

II. Related Work

Machine learning is a type of artificial intelligence (AI) that helps to provides today's computers with the ability to learn without being explicitly programmed by the programmer. Machine learning techniques mainly focuses on the development of computer programs that can teach themselves to grow and change when exposed to new data. Machine learning technology is currently well suited for analyzing medical data. In practical there is a lot of work done in medical diagnosis in small specialized diagnostic problems.

Craven [3] "Representation of sentence using Hidden Markov Models for information" It mainly focuses on entity recognition for diseases and its treatment, relation discriminination by using Hidden Markov.

Bunesce, R. Mooney [4] here Syntaic rule-base relation extraction was implemented. Syntaic rule-base relation extraction systems are complex systems base on additional tool used to assign postages or to extract syntactic parse tree. It is known that in the biomedical literature such tools are not yet at the state of the art level.

R.Gaizauskas[5]In this study they combined syntactic and Semantic rules from Medline abstract in order to obtain better system with the flexibility of the syntactic information and the good precision of semantic rule.

Rosario et al. [6] introduced the Machine Learning (ML) Approach for Identifying Disease-Treatment Relations in Short Texts. The author's mainly focuses on entity recognition for diseases and treatment, relation discriminination by using Maximum Entropy and Hidden Markov models. There are three major approaches used in extracting relations between entities: Statistical methods, co-occurrences analysis and rule based approaches. The system contains both types of sentences i.e. informative and non informative sentences. So the quick access of reliable information is not possible. This is the major drawback of the existing system. Also it is difficult to classify the sentences because of the eight semantic relations which also lead to confusion.

ArzuBaloglu[7] presented the Web Blog Mining Application for Classification of Movie Reviews. In this study, the author's introduced an architecture view, implementation, and evaluation of a Web blog mining application, called as the Blog Miner. This is used to classify and extract people's opinions and emotions (or sentiment) from the contents of weblogs about movie reviews. Also they introduced an opinion mining application that is created for calculating movie scores from Web blog pages. Experimental results show that the proposed application produces accurate results close to real values using key word algorithm. In this study, they introduced only unsupervised approach for sentiment analysis.

MordechaiAverbuch[8] introduced the Context-Sensitive Medical Information retrieval. This paper proposed new algorithms such as learning algorithm and retrieval algorithm for identifying and selecting context in free-text medical narratives. Sometimes the boundary of a sentence cannot be specified correctly. In this case, one sentence is broken into two sentences, or two sentences are considered as one sentence. This is the major drawback of learning algorithm. The retrieval algorithm did not detect the negation and mistakenly retrieved the document. This is the major problem of Retrieval algorithm.

III. Methodology

In this section we have explained the Techniques used for extraction of best Treatment for diseases from short text. The main aim of the system is to generate a Health information recording and clinical data repositories — which will give the doctor an immediate access to patient diagnoses, allergies and lab test results that helps for better and time-efficient medical decisions; Medication management—rapid access to information regarding potential adverse drug reactions, immunizations, supplies, etc; Decision making is the ability to capture and use quality medical data for decisions in the workflow of healthcare.

3.1 Task and Data Set

The two tasks that are undertaken here provide the basis for the design of an information technology framework that is capable to identify and disseminate healthcare information. The first task is identification and extractions of informative sentences on diseases and related treatments topics, while the second one performs a finer classification of these sentences according to the semantic relations that exists between diseases and treatments. The data set consists of sentences from Medline5 abstracts annotated with disease and treatment entities and with eight semantic relations between diseases and treatments. The numbers in parentheses represent the training and test set size. For example, for Only DIS relation, out of 616 sentences present in the data set, 492 are used for training and 124 for testing [7].

There are at least two challenges that can be encountered while working with ML techniques. One is to find the most suitable model for prediction purpose. The ML field also offers a suite of predictive models (algorithms) that can be used and deployed. The task of finding the best suitable one relies heavily on empirical studies and knowledge expertise. The second challenge arised is to find a good data representation technique and to do feature engineering on it because features strongly reflect the performance of the models. Further identifying the correct and sufficient features to represent the data for the prediction models, when the source of information avaliable is not too large, as it is the case of sentences, is an important aspect that needs to be taken into consideration. It very important to identify the sentences extracted is informative or non-informative and labels them so that they can be automatically ignored and the load on the system being developed will be less.

Table 1. Example of Annotated Sentences for sentence Selection task [6]

Label	Sentences
Informative Sentences	Urgent Colonoscopy for the diagnosis and treatment of server diverticular hemorrhage.
NON-informative Sentences	In all cases a coproparasitological study was performed.

3.2 Classification Algorithm

In Machine Learning approach the expertise and previous research provides the guidance to solve new tasks. The models described should be able to identify and provide informative sentences and relation between entities. The research should be made in a way to achieve high performance. As classification algorithms, set of six representative models can be used. They are: adaptive learning (Ada-Boost), decision-based models (Decision trees), and probabilistic models (Naive Bayes (NB) and Complement Naive Bayes (CNB), which is adapted for text with imbalanced class distribution), a linear classifier (support vector machine (SVM) with polynomial kernel), and a classifier that predicts the majority class in the training data. These classifiers are used to learn more algorithms and to work on long text and short texts. Probabilistic models based on Naive Bayes used in text classification and automatic text classification tasks. Decision trees based on decision models are used in short text. Adaptive learning algorithm is used to mainly focus on hard concepts such as unbalanced data sets, unrepresented in data available. [7]

3.3 Data Representation:

3.3.1 Bag-of-Words model [8]

The bag-of-words model is a simplifying representation used in natural language processing and information retrieval. In this model, a text (such as a sentence or a document) is represented as an unordered collection of words and even word order. Recently, the model of bag-of-word has also been used for computer vision.[8]

The bag-of-words model is used in some document classification methods. When a naïve Bayes classifier is applied to text document, for example, the conditional independence assumption adds the assumption that terms are conditionally independent given the class.[8] Other methods of document classification that use this model are latent Dirichlet allocation and latent semantic analysis.

Example:

Text document representation based on the Bag-of-Words model. Here are two simple text documents given below:

Pooja likes to play in rain. Raj likes too.

Pooja also likes to watch TV cartoons.

Based on these two text documents, a dictionary is constructed as shown below:

{"Pooja": 1, "likes": 2, "to": 3, "play": 4,”in”: 5, "rain": 6, "also": 7, "watch": 8, "TV": 9, “cartoons”: 10,"Raj": 11, "too": 12}

This has 12 distinct words. And using the indexes of the dictionary created, each document is represented by a 10-entry vector as shown below:

[1, 2, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1]

[1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 0]

Where each entry of the vectors refers to count of the corresponding entry in the dictionary (this is also the histogram representation). This vector representation does not help to preserve the order of the words in the original sentences. This type of representation has several very successful applications, for example email filtering.

3.3.2. NLP and Biomedical concepts representation [8]

The main developments in this area have been related to the identification of biological entities (named entity recognition), like protein and gene names in free text, the association of gene clusters obtained by microarray experiments with the biological context provided by the corresponding literature, automatic extraction of protein interactions and associations of proteins to functional concepts (e.g. gene ontology terms). Even the extraction of kinetic parameters from text or the sub cellular location of proteins has been addressed by information extraction and text mining technology.

IV. Proposed System

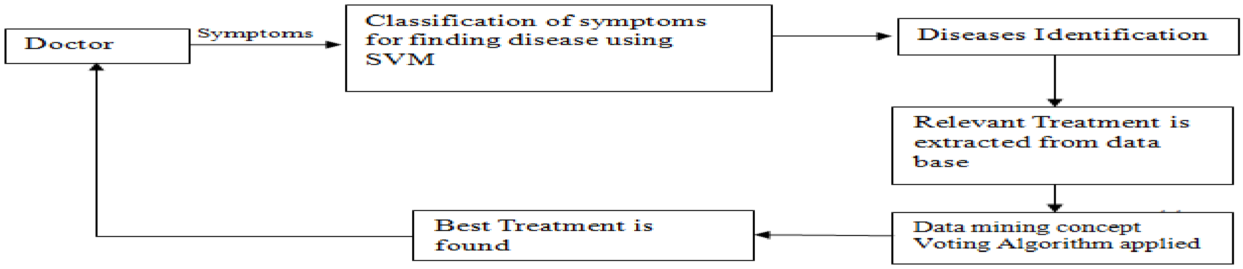
Doctor
Symptoms
Classification of symptoms for finding disease using SVM
Diseases Identification
Relevant Treatment is extracted from data base
Data mining concept Voting Algorithm applied
Best Treatment is found

Fig.1 Architecture of Proposed system for uploading pdf

The above shown is the Fig. 1 architecture of Proposed system for uploading pdf here when doctor uploads a pdf disease file it first verifies whether the pdf is of required format or not if yes it sends the file for preprocessing where the stop word his process is removal, tokenization, bag of words and biomedical concepts are used after this process is over required data is extracted and stored in the database.

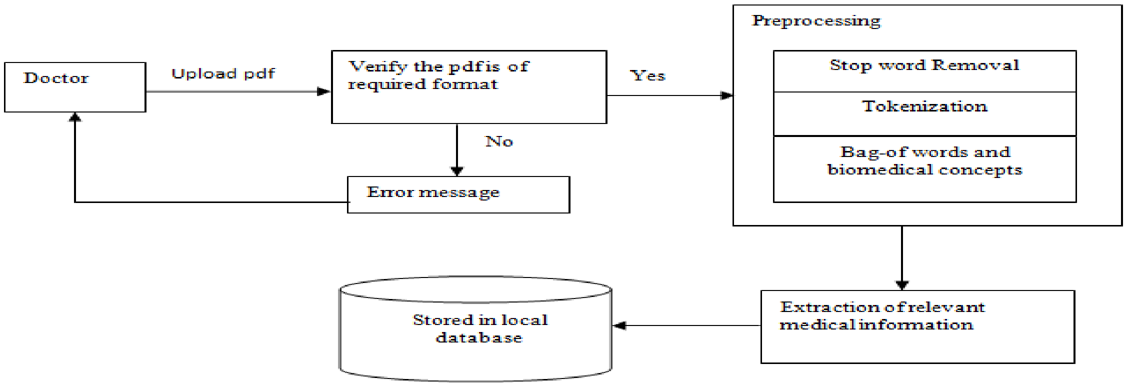
Doctor
Upload pdf
Verify the pdfis of required format
Yes
No
Error message
Preprocessing
Stop word Removal
Tokenization
Bag-of words and biomedical concepts
Extraction of relevant medical information
Stored in local database

Fig. 2 Architecture of Proposed system for treatment extraction

In the above given Fig.2 is architecture of proposed system for treatment extraction on basses of symptoms provided. Here when the doctor enters symptoms using SVM classifier classification of symptoms is done and relevant disease are extracted and out of no. of disease found required disease is selected and which provides with relevant treatment extraction from the database .then using Data mining concept like voting algorithm best treatment is found.

4.1 Input Dataset

In this work, two databases are used. One is Medline database and the other is Local database. The Medline database available contains more than 21 million records from approximately 4,000 selected publications covering biomedicine and health from 1950 to the present. MEDLINE uses Medical Subject Headings (MeSH) for information retrieval. Approximately 4,500 biomedical journals are indexed in MEDLINE. The local database created contains the list of symptoms with the corresponding diseases.

4.2 Data Preprocessing

Data preprocessing describes any type of processing performed on raw data to prepare it for another processing procedure. Commonly used for preliminary data mining practice and also the data preprocessing transforms the data into a format that will be more easily and effectively processed. Initially, the process of splitting the sentence with space using string tokenizer class is done. Then the stop words like a, an, is, was etc are removed. After eliminating the human errors, unwanted words like filler words were removed. Followed by that, stemming is done, which is the process of removing morphological and in flexional ending words to their root words. Finally the semantic word extraction is performed and it is stored in the local database. The same preprocessing techniques such as stemming, stop words removal are performed in Medline database articles.

4.3 Disease Identification

Now the semantic keyword which is a preprocessed symptom is matched with the diseases stored in the local database to identify the corresponding disease related to those symptoms given by the user.

4.4 Extraction of relevant information

After the disease is identified, then articles related to that disease are extracted from medical database available. Then the extracted articles are further classified into informative sentences which contain relevant keywords and non informative sentences which contain irrelevant keywords.

Extract the input articles and then preprocess all the extracted input articles. Then first split the paragraph into sentences using delimiter. Then split each sentence into word using the Stanford POS tagger tool, which also creates tags and words are enclosed with these tags. Obtain the meaning of each word (using parts of speech). Retrieve the informative sentences using the relevant semantic keywords. The advantage is that if a feature appears more than once in a sentence, this means that it is important and the frequency value representation will capture the feature's value will be greater than that of other features.

4.5 Classification

Classification is a data mining (supervised learning) technique used to predict group membership for data instances. Once the related informative sentences are extracted using the keyword searching algorithm, SVM Classifier is used to classify the semantic relations that exists between disease and treatment among the extracted input articles.

Further, data mining (voting algorithm) concept are applied to find the best treatment for Disease. In future, comparative analysis can be done with other classification algorithms in order to provide better performance.

Implementation

The system consist of many medical published articles which can be taken as input for the system and perform our classification and data extraction task on these pdf's.

Algorithms:

```
Step 1: Start
Step 2: Registered Doctor login with his id.
Step 3: Doctor goes to the upload section forstoring data in the local data base.
Step 4: Select a file from medical data
Step 5: Verify whether the file is in required format or not
Step 6: If correct format go to step 7 or stop
Step 7: Read pdf and convert it into text format
Step 8: Then perform preprocessing techniques on it
Use keyword matching algorithm Ai [i=1toN]
Step 9: Extract the input articles.
Step 10: Preprocess all the extracted input articles.
Step 11: First split the paragraph into sentences using delimiter.
Step 12: Next, split each sentence into word using the Stanford POS tagger tool Which also tags and words are enclosed with tags.
Step 13: Obtain the meaning of each word (using parts of speech).
Step 14: Retrieve the informative sentences using the relevant semantic keywordsand Store in local data base
Step15: Doctor enters symptoms as input
Step16: Classification using SVM Classifier is done these symptoms. The classifier constructed as follows
j(i) = no.of disease having symptoms with combinations
cnt= Total no. of diseases
cntsymfound= no. of disease having either 1 or all symptoms
rcount = no. of pdf taken as input.
k(i)=j(i)/cnt;
m(i) = cntsymfound/j(i);
w= (maxsym * ∑y)/rcount
Step 17: Disease related to the symptoms are classified
Step18: The required data related to disease is  extracted from local database
Step19:  Use Voting algorithm for finding best treatment
In the first pass, we need 2 values:
A candidate value initially set to any value
A count initially set to zero
First compare the element's value to the current candidate value.
If they are the same, weincrement count by 1.
If they are different, we decrement count by 1.
The second pass simply counts the frequency of that value to confirm.
Step 20: Best treatment is selected on the bases of majority
Stop 21: Stop.
```

4.6 Performance evaluation

The most commonly used evaluation measures in the machine learning based setting are Accuracy, Recall Precision and F-measure

Accuracy = Total number of correctly classified Instance

$$\text{Recall} = \frac{\text{Correctly classified positive instance}}{\text{total no of correctly classified Instances.}}$$

$$\text{Precision} = \frac{\text{Correctly classified positive instance}}{\text{the total number of positive}}$$

F-measures = The harmonic mean between Precision and recall

V. Results & Discussion

The results of all the above used techniques and algorithms are taken into the account and stored for further calculation. Probabilistic models are stable and reliable for tasks performed on short texts in the medical domain.

The below given comparison table Table 2. tells us, that using SVM Classifier a good result of F-measures= 98.72%, SVM result of Accuracy= 94.80 % is obtained, so in the implementation part using different data representation technique with SVM has obtain a better result. Also the task tackled in our research is to perform with all the above mentioned representations, plus combinations of them. The combination of BOW, NLP and biomedical concepts by putting all features together to represent an instance**.**The results show that probabilistic models based on Naive Bayes formula, obtain good results but the fact that the SVM classifier performs well shows that the current discoveries are in line with the literature.

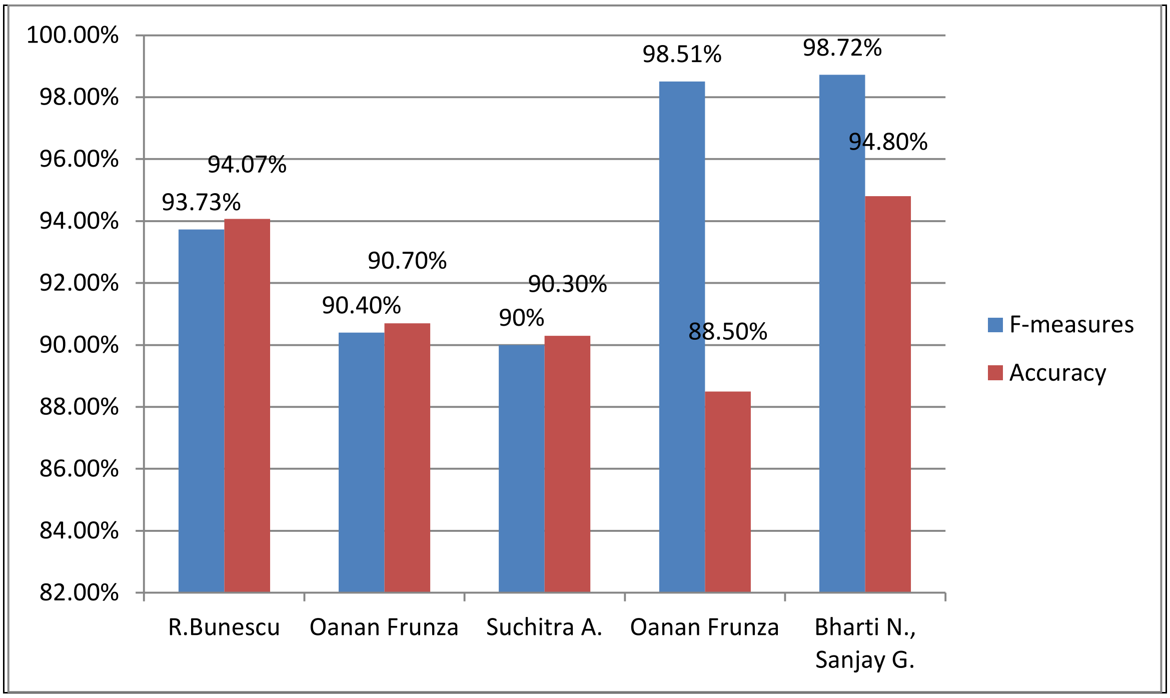

Fig 3. Graphical representation of result and its comparison

Table 2. Comparison of different techniques of machine learning approach in Medical Field and there corresponding

Sr. No.	Paper	Tech. Used	Classification Algorithm	Data Representation	Result
1.	R. Bunescu[8]	Pattern-based method Statistical learning method	SVM classifier Supervised learning Lexical Features Morphosyntactic Features Semantic Features	Bag-of-words representation NLP and biomedical concepts representation Medical concepts (UMLS) representation	Result obtained Accuracy =93.73% F-measures=94.07%
2.	OanaFrunza[10]	The sentence identifies The relation identification based on NLP and ML techniques	Decision-based models (Decision trees) Probabilistic models (Naive Bayes (NB) Complement Naive Bayes (CNB) Imbalanced class distribution Adaptive learning (Ada- Boost) A linear classifier (support vector machine (SVM)with polynomial kernel) A classifier that always predicts the majority class in the training data (used as a baseline).	The bag-of-words (BOW). Concept Type Con Text Verb phrases Concepts Semantic vectors.	F-measures= 90.4% Accuracy=90.7%
3.	Suchitra A[11]	Co-occurrences analysis Rule based approaches Statistical models Inductive Logic Techniques Support Vector Machine Bloom filter is used for the removal of unwanted words so as to fetch only the important words	Same as above	Bag-of-words representation NLP and biomedical concepts representation Medical concepts (UMLS) representation	F-measure=90% Accuracy=90.3%
4.	Oana Frunza [12]	Same as above	Decision-based models (Decision trees) Probabilistic models (Naive Bayes (NB) Complement Naive Bayes (CNB) Imbalanced class distribution Adaptive learning (Ada- Boost) A linear classifier (support vector machine (SVM)with polynomial kernel) A Zero classifier that always predicts the majority class in the training data (used as a baseline	Bag-of-words representation NLP and biomedical concepts representation Medical concepts (UMLS) representation MetaMap is a tool created by NLM that maps free text to medical concepts used in the UMLS	For the relation Cure the F-measure baseline is 98.51%, for Prevent and Side Effect 88.5%
5.	Oana Frunza[13]	The bag-of-words (BOW). Concept Type Con Text Verb phrases Concepts Semantic vectors.	SVM implementation with polynomial kernel from the Weka5 tool.	The bag- of-words (BOW). Concept Type Con Text Verb phrases Concepts Semantic Vectors.	F-measure=86.15% Accuracy =83.5%
6.	Bharti N., Sanjay G.	Bag-of words And Support Vector Machine	Support Vector Machine with linear classifier and Voting Algorithm.	Bag-of words Tokenization Biomedical concepts Stop word removal	F-measures=98.72% Accuracy= 94.80%

VI. Conclusion

We have discussed number of techniques for identification of treatment on disease given. The proposed approach relies on Machine learning and NLP techniques i.e. (i) a pattern-based technique and (ii) a supervised learning method with an SVM classifiers. Compared it to the methods applied separately. The obtained results show that the applied approach significantly outperforms the two mentioned techniques and provides a good alternative to enhance machine learning performance. In this proposed work, keyword searching algorithm used to retrieve relevant healthcare information for the corresponding user symptoms and the classifier are used to classify the semantic relations between disease and treatment. Data mining concept are applied to find the best treatment for Disease.

As future work, different formats of pdf's could be accepted by the system. Also will try to focus more on adding features that are specific for each concept, and to reduce the context from sentence level to shorter contexts, look into more verb information, and better understanding and incorporate additional information for each relation. Further comparative analysis can be done with other classification algorithms in order to provide better performance.

References

[1] Wernick, Yang, Brankov, Yourganov and Strother, "Machine Learning in Medical Imaging", IEEE Signal Processing Magazine, vol. 27, July 2010, pp. 25-38

[2] Agrawal. R, Imieliński T and Swami A "Mining association rules between sets of items in large databases", International Conference on Management of data, 1993, pp. 207-216.

[3] S. Ray and M. Craven, "Representing Sentence Structure in Hidden Markov Models for Information Extraction," Proc. International Joint Conférence Artificial Intelligence, 2001.

[4] R. Bunescu and R. Mooney, "A Shortest Path Dependency Kernel for Relation Extraction," Human Language Technology and Empirical Methods in Natural Language Processing 2005, pp. 724-731.

[5] Srinivasan P. and T. Rindflesch, "Exploring text mining from Medline," Proceedings of the AMIA Symposium, 2002, pp.722–726.

[6] OanaFrunza, Diana Inkpen, and Thomas Tran, "A Machine Learning Approach for Identifying Disease-Treatment Relations in Short Texts" IEEE Transactions On Knowledge And Data Engineering, June 2011, pp. 801-811.

[7] Arzu Baloglu and Mehmet S. Aktas, "Web Blog Mining Application for Classification of Movie Reviews", 5th International Conference on Internet and Web Applications and Services (2010).

[8] Mordechai Averbuch, Tom, Karson "Context-Sensitive Medical Information Retrieval" Health Technology and Informatics, 2004, pp. 282-286.

[9] Rosario and M.A. Hearst, "Classifying Semantic Relations in Bioscience Text," Meeting on Assoc. for Computational Linguistics, vol. 430, 2004.

[10] Razvan c. Bunesc, " Subsequence Kernels for Relation Extraction", Advances in Neural Information Processing Systems, Vol. 18: Proceedings of the 2005 Conference,

[11] Claudio Giuliano and Alberto Lavelli and Lorenza Romano, "Exploiting Shallow Linguistic Information for Relation Extraction from Biomedical Literature",Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics EACL 2006, Trento, Italy, Apr 3, 2006,pp.401-408.

[12] Suykens, Johan A. K.; Vandewalle, Joos P. L."Least squares support vector machine classifiers," Neural Processing Letters, vol. 9, no. 3, Jun. 1999, pp. 293–300.

[13] Zhu Zhang, Xin Li, and Hsinchun Chen "Kernel-Based Learning for Biomedical Relation Extraction", Journal Of The American Society For Information Science And Technology, 2008, pp. 756–769

[14] Boser, Bernhard E, Guyon, Isabelle M, and Vapnik, "A training algorithm for optimal margin classifiers", 5th Annual ACM Workshop on COLT, 1992, pp. 144–152.

[15] Suchitra A and Sudha R "Extraction of Semantic Biomedical Relations from Medline Abstracts using Machine Learning Approach" National Conference on Advances in Computer Science and Applications with International Journal of Computer Applications 2012.

[16] Oana Frunza and Diana Inkpen " Extraction of Disease-Treatment Semantic Relations from Biomedical Sentences" Workshop on Biomedical Natural Language Processing, July 2010,pp. 91-98.

[17] Oana Frunza and Diana Inkpen"Extracting Relations between Diseases, Treatments, and Tests from Clinical Data**"** Canadian conference on Advances in artificial intelligence, 2011, pp. 140–145.

[18] P. Menaka and Prof.D.Thilagavathy "Identifying Semantic Relations for Disease - Treatment in Medline" International Journal of Electronics and Computer Science Engineering, 2012, pp 566-571.

[19] Mr. P. Bhaskar and Mr. Dr. E. Madhusudhana Reddy "Efficient Machine Learning Approach for identifying Disease-Treatment Semantic Relations from Bio-Medical Sentences," International Journal Of Computational Engineering Research (ijceronline.com) Vol. 2 Issue. 5, September 2012, pp. 1425-1429.

[20] Razik Khan, MayurDhande, AniketPatil, NamrataGaikwad "To Identify Disease Treatment Relationship in Short Text Using Machine Learning & Natural Language Processing", Jounarl of Engineering, Computer and applied Science(JEC&AS) VOL.2 Apirl 2013, pp. 72-75.

[21] Y. H. LI and A. K. JAIN, "Classification of Text Documents", THE COMPUTER JOURNAL, Vol. 41, No. 8, 1998, pp.537-546.