# Classification of Diseases and their Treatments Using Machine Learning Approach

**Aditi A Ghive**

**D. R. Patil**

*Abstract* -**In this age of science and technology life has become so hectic so everyone needs to pay attention towards heath care. On the same time everyone needs online heath care system which can help common people to identify their diseases and find treatment for diseases. This will not only help common person but also to doctors to update their knowledge and have correct treatment of diseases. The medical field is one of the fields in which new research is carried out at a faster rate. In a Medical field automation is gaining momentum. From the medical data useful information can be extracted and made useful for generating software's or MEDLINE applications that can help doctors in the treatment. This paper presents a classification of diseases and their treatments using processed biomedical data for classifying diseases using SVM and NAÏVE BAYES classification algorithm. The experimental result shows that SVM gives better Classification Rate than NAVIBAYSE.**

*Keywords:* **MEDLINE, SVM, NAÏVE BAYES, NLP**

## I. INTRODUCTION

In this era life is hectic and due to busy schedules people want each and everything to go in a good flow. Everyone cares for health and wants to be always fit and good health. People want quick access to reliable information. As all are busy so they need to complete their day to day activities quickly using smart technology. One such activity is to look after health. The traditional healthcare system involves long duration so in order to save money and time it needs to be modernized. Diagnosis of various diseases is now carried out by advance healthcare system which involves basic features such as gathering clinical information which has been unutilized from a long period for extraction of useful data which can be used for identification of diseases and finding relations for treating various diseases. Also the research in medical domain and pharmaceutical field can be made available to everyone. Recent developments of drugs on various diseases and evaluation of new diseases and all information related to it can easily available to doctors over the globe. Some diseases are new to the doctors which are not studied during his curriculum which can be easily known and understood by him to treat particular patient. So treating patients with new unknown diseases will be possible by developing a computer application called electronic healthcare system which can use research abstracts throughout the world and finds a relation of disease for proper treatment. This application will be an important approach for modernization of our traditional health care system. Both doctors and patients are benefitted using this EHS (electronic healthcare system). A machine learning technique is an application that is capable of automated identification and dissemination of healthcare information. It extracts sentences from published medical papers that mention diseases and treatments and identifies semantic relations that exist between diseases and treatments.

This paper has introduced natural language processing and machine learning techniques. SVM and NB algorithms are used to identify and classify the medical information in short texts [1].

## II. RELATED WORK

R.Bunescu et al. have used pattern-based method and statistical learning method for classifying diseases and treatments using SVM classifier.ACE corpus (NIST) Dataset have used. The training Part of this dataset consists of 422 documents, with a separate set of 97 documents allocated for testing. Result shows accuracy is 93.73% and f-measure is 94.07% [1].

M.Craven et al. have used dataset for YPD database. He collected 1,213 instances of the sub cellular-localization relation that are asserted in the YPDd database and from MEDLINE. It have used method application domain is novel and challenging; Investigate an approach to decreasing the cost of learning information-extraction routines. Results show Naive Bayes classifier trained on the YPD data reaches 77% precision at 30% recall [2].

A. Suchitra et al. have used three methods co-occurrences analysis rule based approaches, statistical models, and inductive logic techniques. bloom filter is used for the removal of unwanted words so as to fetch only the important words. Probabilistic NB, complement NB and SVM algorithms are used. Results show accuracy 90% and f-measure 90.3% [3].

O. Frunza et al. have used method for bag-of-word, Concept Type Verb phrases Concepts Semantic vectors. SVM classification algorithm is used for information extraction. A result shows F-measure for the *86.3%* and Accuracy 83% [4].

O. Frunza et al. have used method for Decision based model, NAVIBAYES, Ada-Boost. It has used H.Rosario Dataset 2004. Their evaluation result shows 98.55% F-measure for the *Cure* relation, 100% F-measure for the *Prevent* relation, and 88.89% F-measure for the *Side Effect* relation [5].

Rosario et al. have introduce three major approaches for extracting relations between entities: co-occurrences analysis, rule based approaches and Statistical methods. The system contains informative as well as non informative sentences. It have used SVM and NAVIBAYES algorithm. Hearst Rosario Dataset 2004 have used. Result shows 100%F-measure for the Cure relation, 100 % F-measure for Prevent relation, and 75 %F-measure for Side Effect [6].

## III. METHODOLOGY

The basic three Stages to accomplish the objective of classify and identify disease and theirs treatments:-

### A. Preprocessing

Data preprocessing describes any type of processing performed on raw data to prepare it for another processing procedure. Commonly used as a preliminary data mining practice, data preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the user. In this phase noisy and irrelevant data are removed from the extracted data.Initially,the process of splitting the sentence with space using string tokenizer class is done. Then the stop words like a, an, the, 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. The same preprocessing techniques such as stemming, stop words removal are performed in Medline database articles[4].

### B. Classification Algorithms

Classification is a data mining (supervised learning) technique used to predict group membership for data instances. Once the informative sentences are extracted classifiers are used to classify the semantic relations that exists between disease and treatment among the extracted input articles[3].

The role of this system is to identify informative sentences and discriminate disease and its treatment based on semantic relations using ML techniques. We use probabilistic models (Naive Bayes (NB)) and a linear classifier (support vector machine), and a classifier that always predicts the majority class in the training data. These classifiers are used to work on long text and short texts and to learn more algorithms. Probabilistic models are used in automatic text classification tasks [6].

#### a. Naive Bayes

The Naive Bayesian classifier is based on Bayes' theorem with independence assumptions between predictors. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets [2].

Bayes theorem provides a way of calculating the posterior probability, P (c|x), from P(c), P(x), and P (x|c). Naive Bayes classifier assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence [5].

Likelihood Class Prior Probability

$$P(c|x)=\frac{P(x|c)p(c)}{P(x)} \qquad (1)$$

Posterior Probability Predictor Prior Probability

$$P(c|x)=P(x_1|c)\times P(x_2|c)\times\ldots.\times P(x_n|c)\times P(c) \qquad (2)$$

P (c|x) is the posterior probability of class (target) given predictor (attribute).

P (c) is the prior probability of class.

P (x|c) is the likelihood which is the probability of predictor given class.

P (x) is the prior probability of predictor.

#### b. SVM

In machine learning, support vector machines are supervised learning models with associated learning algorithm that analyze data and recognize patterns, used for classification and regression analysis [1]. Given a set of training examples, each marked as belonging to one of two categories, a SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. A SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on. The idea for SVM is to find a boundary. SVM does this by taking a set of points and separating those points using mathematical formulas[5].

The outline of our algorithm is presented:-

**SVM Algorithm Steps:-**

```
1. Simple SVM
2. CandidateSV = { closest pair from opposite classes}
   **while** there are violating points **do**
3. Find a violator
4. CandidateSV = candidateSV
   S
  violator
  **if** any _p < 0 due to addition of c to S **then**
5. candidateSV = candidateSV \ p
   repeat till all such points are pruned
**end if**
   **end while**
```

### C. Dataset

Rosario and Hearst (2004) dataset have used to entire data set is collected from Medline2 2001 abstracts. Sentences from titles and abstracts are annotated with entities and with 8 relations, based only on the information present in a certain sentence. The first 100 titles and 40 abstracts from each of the 59 Medline 2001 files were used for annotation [6].

#### Relationships

This section describes the various types of semantic relations that were found to occur between the semantic classes of treatment and disease. Below are shown a few examples for each relationship.

#### Cure

To label a sentence as `cure', the treatment has to cure the disease <label> means that the word that follows it is the first of the entity and </label> that the word that proceeds it is the last of the entity. Examples for this relation are <DIS> Obesity </DIS> is an important clinical problem and the use of <TREAT> dexfenfluramine hydrochloride /TREAT for weight reduction has been widely publicized since its approval by the Food and Drug Administration.

**Only Disease**

When a treatment was not mentioned in the sentence (other entities may have been present). Some examples:

The objective of this study was to determine if the rate of <DISONLY> preeclampsia </DISONLY> is increased in triplet as compared to twin gestations.

**Only Treatment**

When a disease was not mentioned in the sentence (other entities may have been present). Some examples:

Patients were randomly assigned either <TREATONLY> roxithromycin </TREATONLY> 150 mg orally twice a day

**Prevent**

When there is a clear implication that a <TREAT> will prevent a <DIS>. This might be inherent in the definition of the treatment, e.g. a vaccine works by preventing a disease from occurring, or explicitly stated, often with the words ``prevent'' or ``prevention of''. Also seen is the phrase ``reduce incidents'', ``reduce rates of'', or ``reduction in rates...'' because these also imply that disease events are being prevented. Examples:

Immunogenicity of <DIS PREV> hepatitis B </DIS PREV> <TREAT PREV> vaccine </TREAT PREV> in term and preterm infants.

**Side Effect**

When a DISEASE is a result of a TREATMENT. The cause/effect relationship should be explicitly stated or at least very clearly implied. The most common toxicity is <DIS SIDE EFF> bone pain </DIS SIDE EFF>, and other reactions such as <DIS SIDE EFF> inflammation </DIS SIDE EFF> at the site of <TREAT SIDE EFF> injection </TREAT SIDE EFF> have also occurred.

**Vague**

When there is semantically a very unclear relationship between a TREATMENT and a DISEASE. It can be either a TREATMENT that affects a DISEASE or something associated with the condition of a DISEASE or, not as often, a DISEASE that has some sort of effect on a TREATMENT.

Example: <TREAT VAG> Hormone replacement therapy </TREAT VAG> and <DIS VAG> breast cancer </DIS VAG>

## IV. Experimental Results

The implementation has been tested on MATLAB-R2013a, with system having Intel Core i7 2630QM Processor 2GHz, 8GB DDR3 RAM, HD Graphics 3000, Windows 7.1. Disease relationship used here are from Hearst Rosario Data Set dataset have used to entire data set is collected from Medline2 2001 abstract[6].

F-measure is calculated in both algorithms. The F-Measure computes some average of the information retrieval precision and recall metrics. An arithmetic mean does not capture the fact that a (50%, 50%) system is often considered better than an (80%, 20%) system. It shows TP, FP, TN, and FN are the number of true/false positives/ negatives [3].

- Precision: p=TP/(TP+FP)
- Recall: r= TP/(TP+FN)

F-measure is computed using the harmonic mean:

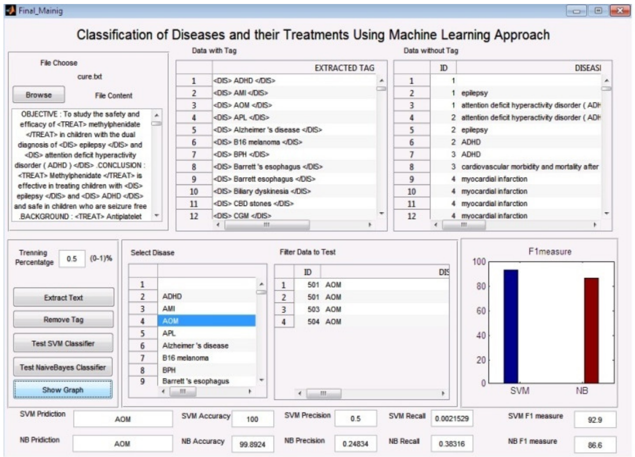
Given n points, x1, x2, xn, the harmonic mean is

$$\frac{1}{H}=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{X_i} \qquad (3)$$

So, the harmonic mean of Precision and Recall:

$$\frac{1}{F}=\frac{1}{2}\left(\frac{1}{R}+\frac{1}{P}\right)=\frac{P+R}{2PR} \qquad (4)$$

The following system show the working of the model is tested with the text file from MEDLINE containing information about all disease and theirs treatments. It can also calculate the value precision, recall, accuracy and f-measure of particular type of diseases.



Classification of Diseases and their Treatments Using Machine Learning Approach

File Choose: cure.txt

File Content

OBJECTIVE : To study the safety and efficacy of <TREAT> methylphenidate </TREAT> in children with the dual diagnosis of <DIS> epilepsy </DIS> and <DIS> attention deficit hyperactivity disorder ( ADHD ) </DIS> .CONCLUSION : <TREAT> Methylphenidate </TREAT> is effective in treating children with <DIS> epilepsy </DIS> and <DIS> ADHD </DIS> and safe in children who are seizure free .BACKGROUND : <TREAT> Antiplatelet

Data with Tag

| | EXTRACTED TAG |
|---|---|
| 1 | <DIS> ADHD </DIS> |
| 2 | <DIS> AMI </DIS> |
| 3 | <DIS> AOM </DIS> |
| 4 | <DIS> APL </DIS> |
| 5 | <DIS> Alzheimer 's disease </DIS> |
| 6 | <DIS> B16 melanoma </DIS> |
| 7 | <DIS> BPH </DIS> |
| 8 | <DIS> Barrett 's esophagus </DIS> |
| 9 | <DIS> Barrett esophagus </DIS> |
| 10 | <DIS> Biliary dyskinesia </DIS> |
| 11 | <DIS> CBD stones </DIS> |
| 12 | <DIS> CGM </DIS> |

Data without Tag

| | ID | DISEASI |
|---|---|---|
| 1 | 1 | |
| 2 | 1 | epilepsy |
| 3 | 1 | attention deficit hyperactivity disorder ( ADH |
| 4 | 2 | attention deficit hyperactivity disorder ( ADH |
| 5 | 2 | epilepsy |
| 6 | 2 | ADHD |
| 7 | 3 | ADHD |
| 8 | 3 | cardiovascular morbidity and mortality after |
| 9 | 4 | myocardial infarction |
| 10 | 4 | myocardial infarction |
| 11 | 4 | myocardial infarction |
| 12 | 4 | myocardial infarction |

Trenning Percentatge: 0.5 (0-1)%

Select Disase

| | |
|---|---|
| 1 | |
| 2 | ADHD |
| 3 | AMI |
| 4 | AOM |
| 5 | APL |
| 6 | Alzheimer 's disease |
| 7 | B16 melanoma |
| 8 | BPH |
| 9 | Barrett 's esophagus |

Filter Data to Test

| | ID | DIS |
|---|---|---|
| 1 | 501 | AOM |
| 2 | 501 | AOM |
| 3 | 503 | AOM |
| 4 | 504 | AOM |

SVM Pridiction: AOM  SVM Accuracy: 100  SVM Precision: 0.5  SVM Recall: 0.0021529  SVM F1 measure: 92.9

NB Pridiction: AOM  NB Accuracy: 99.8924  NB Precision: 0.24834  NB Recall: 0.38316  NB F1 measure: 86.6

Fig.2. **Classification of Diseases and their Treatments Using Machine Learning Approach**

| **Dataset** | Cure | | Cure | | Cure | |
|---|---|---|---|---|---|---|
| **Disease** | AMI | | BPH | | CARBON MONOXIDE POISONING | |
| **Algorithm** | SVM | NB | SVM | NB | SVM | NB |
| **% Accuracy** | 100 | 100 | 100 | 100 | 100 | 100 |
| **Precision** | 0.500 | 0.252 | 0.000 | 0.178 | 0.5000 | 0.315 |
| **Recall** | 0.001 | 0.408 | 0.000 | 0.361 | 0.0013 | 0.442 |
| **F-measure** | 92.90 | 88.90 | 60.70 | 80.90 | 90.30 | 80.00 |

**Table1.Classification Result for SVM and NB**

## CONCLUSION

In this paper MEDLINE dataset taken by Hearst Rosario is processed using NLP and bag-of-words representation techniques. The processed dataset is further utilized to extract keywords such as cure, only disease, only treatment, prevent, side effect, vague, does not cure, complex and none. SVM and Naïve bayes algorithm are used as classifier to classify diseases and their treatments. After comparing results of both algorithms based upon the values of F-measure. The experimental result shows that SVM gives better classification rate result than NAVIBAYSE. SVM gives more accuracy than NB .Greater the value of f-measure more is the accuracy.

## REFERENCES

[1] R.C.Bunescu and R.J. Mooney, "A Shortest Path Dependency Kernel for Relation Extraction", Proc.Conf. Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP), pp. 724-731, 2005.

[2] M. Craven, "Learning to Extract Relations from Medline," Proc. Assoc. for the Advancement of Artificial Intelligence, 1999.

[3] 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.

[4] O. Frunza and D. Inkpen, " Extracting Relations Between Diseases, Treatments, and Tests from Clinical Data", C.Butz and P. Lingras (Eds): Canadian AI 2011, LNAI 6657, Springer-Verlag Berlin Heidelberg, pp. 140-145, 2011.

[5] O. Frunza, D. Inkpen and T. Tran, "A Machine Learning Approach for Identifying Disease Treatment Relations in Short Texts", *IEEE* Transactions on Knowledge and Data *Engineering,* Vol. 23, No. 6, June 2011.

[6] Rosario and Hearst, "Machine Learning (ML) Approach for Identifying Disease -Treatment Relations in Short Texts", *IEEE* Transactions on Knowledge and Data *Engineering, 2011*.

[7] http://biotext.berkeley.edu/dis_treat_data.html

## AUTHOR'S PROFILE

**Aditi Ghive** received the B.E degree in computer science from North Maharashtra University, Jalgaon, in 2011 and currently pursuing M.E degree in computer science from North Maharashtra University, Jalgaon. She has 3 publications in reputed conference. Her research includes machine learning and data mining

**D.R.Patil.** Asst Prof. Dept. of Computer Engineering SES's R. C. Patel Institute of Technology Shirpur