International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 11 (2015) pp. 29005-29014

http://www.ripublication.com

A Naive-Bayes Approach For Disease Diagnosis With Analysis of Disease Type and Symptoms

Shahid Khan[1], Muhammad Rukunuddin Ghalib[2*]
[1,2]School of Computing Science and Engineering
[1]kshahid000@gmail.com, [2]ghalib.it@gmail.com
Vellore Institute of Technology, Vellore, Tamil Nadu, India
**Corresponding Author: Muhammad Rukunuddin Ghalib, ghalib.it@gmail.com*

Abstract

The diagnosis of a medical condition is based on the symptoms, physical examination and medical history of a patient. There have been many cases where treatment for a disease has not been carried out accurately due to lack of proper analysis and ignorance of a number of important factors. To make the essentials of good health practices accessible to everyone, a Naïve Bayes approach for Disease Diagnosis is proposed here. This service aims to help patients and doctors by guiding them to predict possible diseases using the symptoms provided by the user. Naïve Bayes classifier is used for classification of data. The system comprises of a symptom data set which is further categorized as Name, Attributes, and Record data. Based on the ranking of each symptom provided, the diagnosed result suggests the possibility weight of each disease. It also gives the detail of identified diseases including the cure, prevention and the possible treatment. An analysis of Naïve Bayes classification accuracy in the system is also studied.

Keywords: Multinomial Naive Bayes, electronic patient record, confusion matrix, classification accuracy, disease classification.

Introduction

In today's environment, it is clearly seen how healthcare sector is modernized with highly equipped machineries and monitoring system [8]. Hospitals and medical centers today produce a huge amount of complex electronic data [8] [9] [17]. One of the examples of crediting and utilizing this data is Electronic Patient Record (EPR) or Electronic Health Record (EHR) [17]. It is basically a way of seeing the collected patient's medical record through computer interface. There are different versions of EPR used for research purpose. Various Data mining and Data classification methods have been used by the researchers to try to improve the healthcare sector [1-9] [12-13]

[17]. Basically these techniques are classified in supervised and unsupervised category [5] [8] [11] [14] [16]. Unsupervised learning has been more popular in pattern recognition and image processing area [8] [11] [14] [16]. In this paper we are focused in predictive methods (specifically Naive Bayes) which are also known as classification models. A few classification methods used in the medical diagnosis are Decision tree, svm, k-NN, Naïve Bayes etc. K-Nearest Neighbor algorithm is based on the assumption that the members of the same class should be similar [11] [16]. It is suited for data stream and does not build classifier in advance. In decision tree the data is categorized in the form of tree structure. It works fast but has a risk of over fitting the data with the occurrence of an alternative tree. Support Vector Machine is statistical based algorithm which discriminates between the positive and negative members of the given class of n-dimensional vector [2] [3].

It is considered to be one of the most efficient classification algorithms but is more complex and requires more time and memory in the training and classification stage. The Naive Bayes classifier is a straightforward probabilistic classifier which is in view of Bayes theorem with solid and guileless independence assumptions [3] [14]. It is a standout amongst the most essential text classification methods with different applications in email spam discovery, individual email sorting, document categorization, sexually explicit content detection, language detection and sentiment detection. Naïve Bayes classifier is exceptionally productive since it is less computationally concentrated (in both CPU and memory) and it obliges a small amount of training data [14] [16]. Additionally, the training time with Naive Bayes is significantly smaller as compared to other methods.

Related Work

A number of researchers have come up with different methods to diagnose diseases using different data mining and classification techniques. In [4] a previous work Baati, K., Hamdani, T.M., Alimi, and A.M. investigate a Nai□ve Bayes Style Possibilistic Classifier (NBSPC) to make decision from the categorical and subjective medical information. The main focus of the work is to improve the classification accuracy. NBSPC simultaneously relies on the structure of the Nai□ve Bayes classifier as a good classifier for categorical features, and on the possibility theory as an interesting framework to model and fuse subjective medical data. In another work [5] Muhammad, L.A.-N. discussed the experiment that was executed with Naive Bayes technique in order to build predictive model as an artificial diagnose for heart disease based on data set which contains set of parameters that were measured for individuals previously. In another research [15] Karakis, R.; Elektron. ve Bilgisayar Egitimi Bolumu, studied five results of pulmonary function test (PFT) evaluated with machine learning methods and feature selections with test results. Feature selections are performed using Naive Bayes, support vector machine (SVM), linear discriminant analysis (LDA) and k-nearest neighbor classifier (k-NN) methods. In [7] work by Balakrishnan, S.; Narayanaswamy, R.; Savarimuthu, N.; Samikannu, R., a feature selection approach for finding an optimum feature subset that enhances the classification accuracy of Naive Bayes classifier was proposed. The experiments were conducted on the Pima

Indian Diabetes Dataset to assess the effectiveness of their approach. In another work [3] Fathima, S.; Hundewale, N., presents the performance analysis of different data mining techniques to predict the Arboviral disease-Dengue. They have investigated SVM, Naive Bayes Classifier and a classifier which allows identifying small sets of parameters to be used for diagnostic purposes in clinical practice.

Methodology

There are several Naive Bayes Variations such as the Multinomial Naive Bayes, the Binarized Multinomial Naive Bayes and the Bernoulli Naive Bayes [20]. Note that each can deliver completely different results since they use completely different models. Here Multinomial Naive Bayes is used which shows its significance when multiple occurrences of the words matter a lot in the classification problem [20]. Such an example is when we try to perform Topic Classification. This variation, as described by Manning et al [19], estimates the conditional probability of a particular word/term/token given a class as the relative frequency of term t in documents belonging to class c as shown in Eqn (1).

(1)

Thus this variation of Naïve Bayes takes into account the number of occurrences of term t in training documents from class c, including multiple occurrences [20]. Both the training and the testing algorithms are presented below in the form of pseudo code [21].

TRAIN MULTINOMIAL NB(C, D)
1 $V \leftarrow$ EXTRACTVOCABULARY(D)
2 $N \leftarrow$ COUNTDOCS(D)
3 **for each** $c \in C$
4 **do** $N_c \leftarrow$ COUNTDOCSINCLASS(D, c)
5 $prior[c] \leftarrow N_c / N$
6 $text_c \leftarrow$ CONCATENATETEXTOFALLDOCSINCLASS(D, c)
7 **for each** $t \in V$
8 **do** $T_{ct} \leftarrow$ COUNTTOKENSOFTERM($text_c$, t)
9 **for each** $t \in V$
10 **do** $condprob[t][c] \leftarrow \frac{Tct+1}{\sum_{t'}(Tct'+1)}$
11 **return** V, $prior$, $condprob$

APPLY MULTINOMIAL NB(C, V, $prior$, $condprob$, d)
1 $W \leftarrow$ EXTRACTTOKENSFROMDOCS(V, d)
2 **for each** $c \in C$
3 **do** $score[c] \leftarrow \log prior[c]$
4 **for each** $t \in W$
5 **do** $score[c]$ += $\log condprob[t][c]$ **return** $\arg\max_{c \in C} score[c]$

Figure 1: Pseudo code of training and testing algorithms used

Experimental Setup and Design

The steps in implementing the pre processing methodology is given in figure 2. It first starts with feeding the document to the algorithm. Tokenization takes place which allows the algorithm to remove the stop words. After tokenization process, stemming is carried for releasing the final processed data.

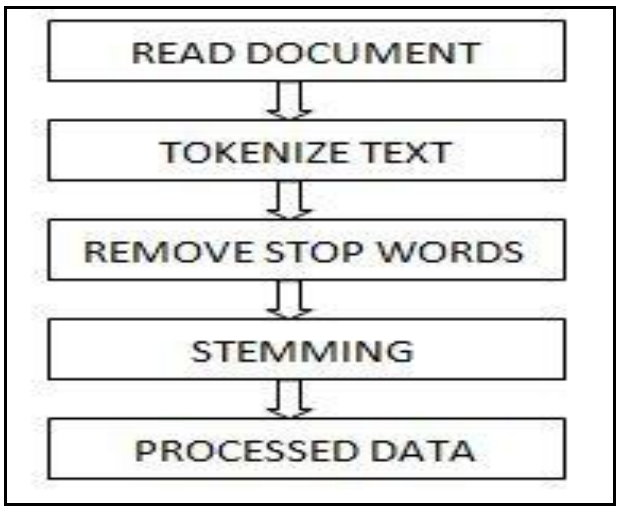

Figure 2: Preprocessing of Data provided to the system

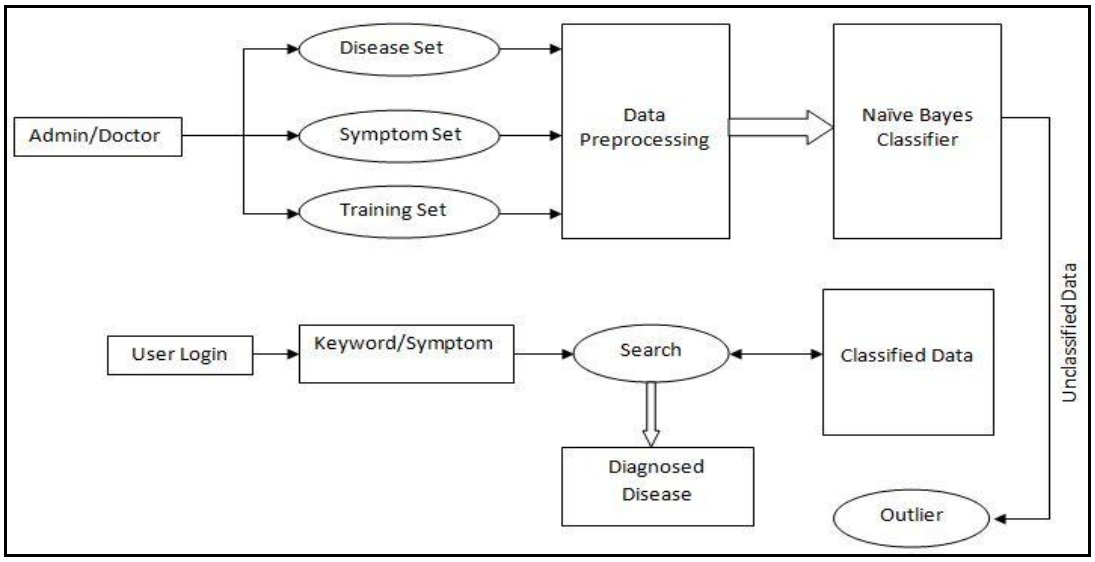

Figure 3: The Proposed Architecture of the Disease Diagnosis system

The implementation of the system starts with the creation of the data sets as shown in figure 3. The data sets consist of Disease data set, Symptom data set and Training data set. These data sets will be uploaded in the system by the administrator, which may be a doctor or any authorized person. The Disease data set contains information

of each disease in the form of 'Prevention', 'Side effects' and 'Treatment'. In the Symptom data set the data is in the form of 'Name', 'Attribute' and 'Data'. The 'Data' in the symptom data set includes the different combination of symptoms taken from old patient records. The combination of symptoms which do not cause the disease is considered to be the outlier (unclassified) as shown in Figure 3. In the Training set the data is trained for two things that is 'Relation Identification' and 'Sentence Selection'. A sample data set of different types of cancer is used to test the proposed system.

Data uploaded in the system is initially preprocessed as shown in figure 1. The preprocessed data is then trained by finding the relation of the symptoms to the diseases. The confusion matrix is calculated for each symptom with all diseases. Figure 4 shows the sample look of the confusion matrix for a particular set of symptoms.

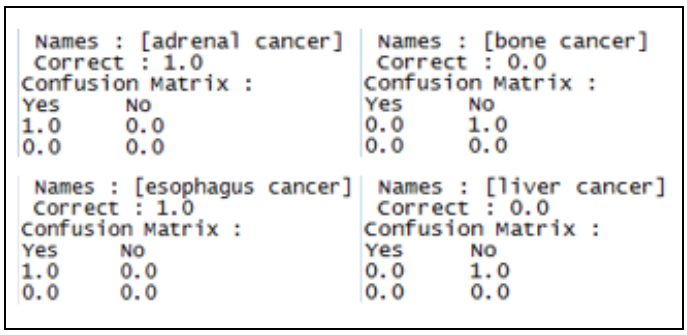

Figure 4: Sample of Confusion Matrix while identifying the diseases

The nominal attribute 'YES OR NO' is used to identify the relation in confusion matrix. Diseases classified as YES in confusion matrix are then processed to find the Informative Sentences (IS) and Non-Informative Sentences (NIS) from available disease data set. Sentences identified as IS are then further processed to be classified in one of the three categories: Cure (C), Prevention (P) or Treatment (T) as shown in Figure 5.

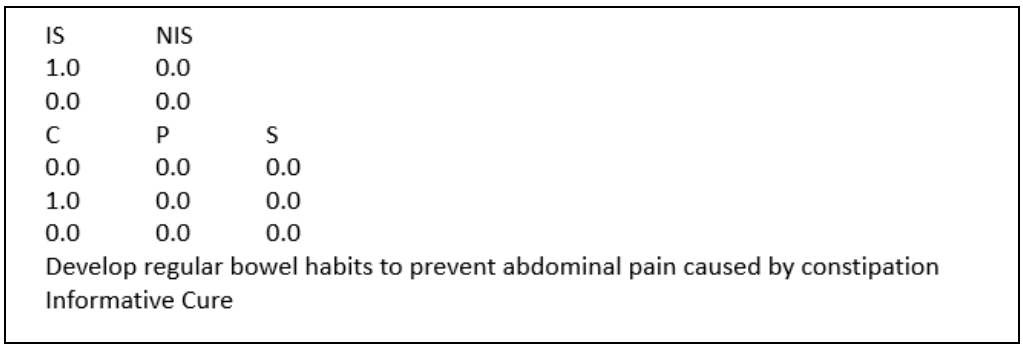

Figure 5: Sample of Sentence Classification wheather IS or NIS and (C) or (P) or (S).

After identifying each category, symptoms are then mapped with the diseases identified. Symptoms provided by the normal user are processed similarly and mapped. Each symptom provided by user is ranked after processing based on the weight and occurrence of that symptom in particular disease. At the end, the total weight of the provided symptoms with each disease is calculated and presented as the possibility weight of the particular disease.

Experimental Result and Discussion

Based on the conducted experiments, the accuracy result for Naïve Bayes with respect to disease data set is presented in Table 1. The combinational symptom data in the data set is made sufficiently available for each nominal attribute. For better results, approximately equal number of data is taken for each nominal attribute. Since the Naïve Bayes classification algorithm works on probability, classes having sufficient data will satisfy the probability condition and be picked up as result. From Table 1 and Table 2 it is analyzed that the classification accuracy varies based on the type of data used [8]. The symptom record in Table 1 shows the available number of record for each nominal attribute (i.e. 40 for each). The outlier accuracy analysis shows that sometimes there is an inaccuracy in the system. This is shown in Table 1 for outlier accuracy in adrenal and eye cancer data set. Table 2 shows the complete analysis of processed and unprocessed sentences in all disease data sets. It clears the amount of IS and NIS sentences processed.

Table 1: Analysis of symptom classification giving the number of each nominal type i.e. 40. The outlier accuracy give number of true classified and false classified. And the percentage of classification

Symptom data set name(cancer)	**No. of symptom attribute**	**Symptom record**	**Nominal (yes & no)**	**Outlier accuracy (yes, no)**	**% of classification accuracy**
Adrenal	14	(40,40)	2	(12,14)	65.00
Bone	7	(40,40)	2	(0,14)	82.50
Esophagus	9	(40,40)	2	(0,16)	80.00
Eye	10	(40,40)	2	(2,16)	77.50
Kidney	7	(40,40)	2	(0,15)	81.75
Liver	11	(40,40)	2	(0,14)	82.50
Lung	15	(40,40)	2	(0,10)	87.50
Stomach	9	(40,40)	2	(0,40)	100.00
Thymus	10	(40,40)	2	(0,12)	85.00
Thyroid	7	(40,40)	2	(0,10)	87.50

Table 2: This is the analysis of disease data set giving the percentage of sentences processed and unprocessed.

Disease data set name	**No. of sentences**	**% of unprocessed**	**% of processed**		
			IS	**NIS**	**Total**
Adrenal	53	22.64	30.18	47.18	77.36
Bone	50	12.00	46.00	41.17	87.17
Esophagus	57	22.80	36.84	40.34	77.18
Eye	56	23.21	30.35	46.42	76.77
Kidney	51	9.80	45.09	35.29	80.38
Liver	57	17.54	36.84	45.61	82.45
Lung	42	11.90	52.38	40.47	92.85
Stomach	82	14.63	50.00	35.36	85.36
Thymus	57	21.05	45.61	33.33	78.94
Thyroid	87	21.83	32.18	45.97	78.15

Classifier Evaluation Measures

To further analyze the performance of the method, few specific evaluation measures [3] [9] [17] are needed to be calculated as shown in Table 3.

Sensitivity is the recognition rate or genuine positive rate while Specificity is the genuine negative rate [9] as shown in Eq (2) and Eq (3).

(2)

(3)

True_Pos is the number of genuine positives (i.e. examples that were effectively classified) and Pos are the quantity of positive specimens [9]. True_Neg is the quantity of genuine negativesand Neg is the number of negative specimens. False positives (F_Pos) are the negative tuples that were inaccurately marked by the classifier [9].

(4)

(5)

The true positives, true negatives, false positives and false negatives are also useful in assessing the costs and benefits (or risks and gains) associated with a classification mode [9].

Table 3: Naïve Bayes Evaluation measures

Accuracy	**Sensitivity**	**Specificity**	**Error Rate**
79	87	45	19

Figure 6 shows the final result of the implemented system. The list given in the Figure 6 are the diagnosed diseases which are possible by the symptoms provided. It also shows the possibility weight of each disease in the diagnosed list. The experimental analysis further gives the limitation of the system that a minimum of 3 symptoms are needed by the user to be given to the system to work correctly.

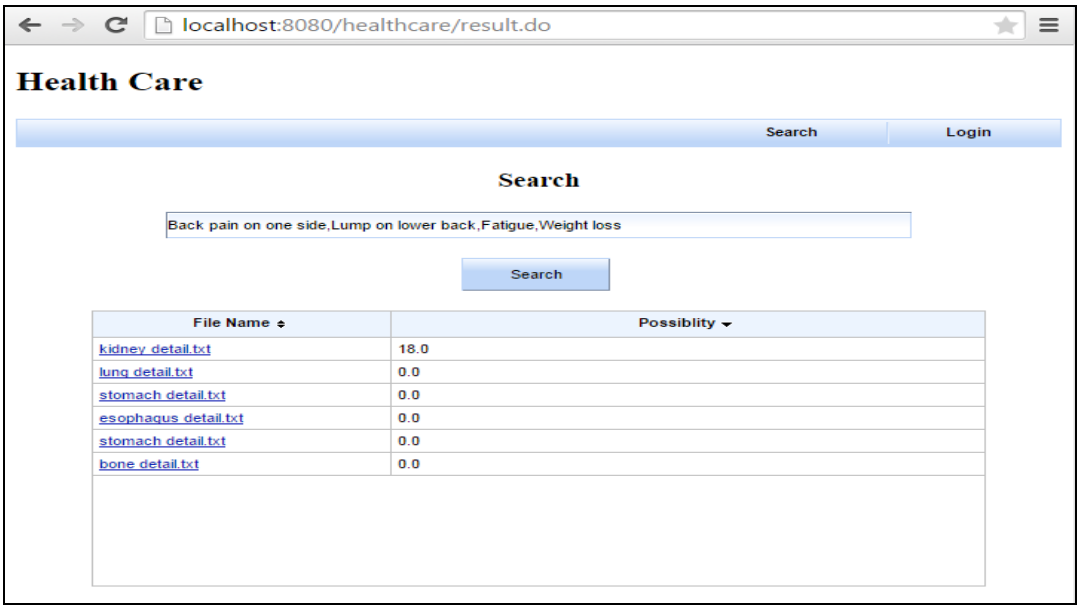

Figure 6: Screen shot of the final result of the system giving the list of diagnosed diseases with is weight of each.

Conclusion and Future Scope

A Disease Diagnosis system is developed using Naïve Bayes Classification technique. The system takes the symptoms as the input from the user and provides the number of possible diseases with those symptoms. The experimental analysis gives a conclusion that the use of Naïve Bayes in the system is efficient enough to be used by patients or doctors to predict the diseases. The detailed information of the system with its model interpretation and accuracy is provided. The system is expandable in the sense that more number of patient record and attributes can be incorporated. Currently only few data sets of cancer is been used in the system. This can be further expanded to other diseases with its symptom data set and disease data set.

Acknowledgement

We express our gratitude to the Department of Computer Science and Engineering, Vellore Institute of Technology, Vellore for providing me with facilities and constant support for carrying out this research.

References

[1] Long Qu; Vetter, V.L.; Bird, G.L.; Haijun Qiu; White, P.S., "A Naïve Bayes classifier for differential diagnosis of Long QT Syndrome in children," *Bioinformatics and Biomedicine (BIBM), 2010 IEEE International Conference on* , vol., no., pp.433,437, 18-21 Dec. 2010.

[2] Jin Huang.; Jingjing Lu.; Charles X. Ling., "Comparing Naive Bayes, Decision Trees, and SVM with AUC and Accuracy", Third 2003 IEEE International Conference on Data Mining (ICDM'03), vol.,no.,pp. 7695,1978 4 Apr.2003.

[3] Fathima, S.; Hundewale, N., "Comparison of classification techniques-SVM and naives bayes to predict the Arboviral disease-Dengue," *Bioinformatics and Biomedicine Workshops (BIBMW), 2011 IEEE International Conference on*, vol., no., pp.538, 539, 12-15 Nov. 2011.

[4] Baati, K.; Hamdani, T.M.; Alimi, A.M., "Diagnosis of Lymphatic Diseases Using a Naive Bayes Style Possibilistic Classifier," *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on* , vol., no., pp.4539,4542, 13-16 Oct. 2013.

[5] Muhammed, L.A.-N., "Using data mining technique to diagnosis heart disease," *Statistics in Science, Business, and Engineering (ICSSBE), 2012 International Conference on* , vol., no., pp.1,3, 10-12 Sept. 2012.

[6] Wagholikar, K.B.; Vijayraghavan, S.; Deshpande, A.W., "Fuzzy naive bayesian model for medical diagnostic decision support," *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, vol., no., pp.3409, 3412, 3-6 Sept. 2009.

[7] Balakrishnan, S.; Narayanaswamy, R.; Savarimuthu, N.; Samikannu, R., "SVM ranking with backward search for feature selection in type II diabetes databases," *Systems, Man and Cybernetics, 2008. SMC 2008. IEEE International Conference on*, vol., no., pp.2628, 2633, 12-15 Oct. 2008.

[8] K.M. Al-Aidaroos, A.A. Bakar and Z. Othman, "Medical Data Classification with Naive Bayes Approach". *Information Technology Journal, 11: 1166-1174.* 2012.

[9] Ms.Rupali R.Patil, " Heart Disease Prediction System using Naïve Bayes and Jelinek-mercer smoothing" *International Journal of Advanced Research in Computer and Communication Engineering,* Vol. 3, Issue 5, May 2014.

[10] S.L. Ting.; W.H. Ip.; Albert H.C. Tsang., "Is Naïve Bayes a Good Classifier for Document Classification?" *International Journal of Software Engineering and Its Applications*, Vol. 5, No. 3, July, 2011.

[11] Baharum Baharudin.; Lam Hong Lee.; Khairullah Khan., *"*A Review of Machine Learning Algorithms for Text-Documents Classification.*"* *Journal of Advances in Information Technology*, Vol 1, No 1 (2010), 4-20, Feb 2010.

[12] Krishnaiah, V., Dr G. Narsimha, and Dr N. Subhash Chandra. "Diagnosis of lung cancer prediction system using data mining classification techniques." *International Journal of Computer Science and Information Technologies* 4.1 (2013): 39-45.

[13] Chaurasia, Vikas and Pal, Saurabh, "Data Mining Approach to Detect Heart Diseases". *International Journal of Advanced Computer Science and Information Technology (IJACSIT)* Vol. 2, No. 4, 2013, Page: 56-66.

[14] Vandana Korde.; C Namrata Mahender., "Text Classification and Classifiers: A Survey" International Journal of Artificial Intelligence & Applications (IJAIA), Vol.3, No.2, March 2012.

[15] Karakis, R.; Guler, I.; Isik, A.H., "Feature selection in pulmonary function test data with machine learning methods," *Signal Processing and Communications Applications Conference (SIU), 2013 21st* , vol., no., pp.1,4, 24-26 April 2013.

[16] Shweta C. Dharmadhikari.; Maya Ingle. Parag Kulkarni., "Empirical Studies on Machine Learning Based Text Classification Algorithms", *Advanced Computing: An International Journal (ACIJ)*, Vol.2, No.6, November 2011.

[17] Mai Shouman.; Tim Turner.; Rob Stocker, "Integrating Decision Tree and K- Means Clustering with Different Initial Centroid Selection Methods in the Diagnosis of Heart Disease Patients", Journal of Computer Science and Engineering, Volume 20, Issue 1, August 2013.

[18] http://www.cs.unb.ca/profs/hzhang/publications/FLAIRS04ZhangH.pdf

[19] http://nlp.stanford.edu/IRbook/html/htmledition/propertiesofnaivebayes1.html

[20] http://blog.datumbox.com/machine-learning-tutorial-the-naive-bayes-text-classifier/

[21] http://blog.datumbox.com/developing-a-naive-bayes-text-classifier-in-java/