Performance Evaluation of Machine Learning Techniques for Text Classification

Dr.R.Manicka chezian
Associate Professor, Department of Computer Science, Nallamuthu Gounder Mahalingam College, Pollachi, India
Email: chezian_r@yahoo.co.in
C.Kanakalakshmi
Research Scholar,Department of Computer Science, Nallamuthu Gounder Mahalingam College, Pollachi, India
Email: kanaks2512@gmail.com

----------**ABSTRACT**----------
Text Mining is the discovery of valuable, hidden information from the text document. Text Classification is the process of classifying documents into predefined categories based on their content. Text classification approach is gaining more importance because of the accessibility of large number of electronic documents from a variety of resources. It is the method of finding interesting regularities in large textual documents. The goal of text mining is to enable users to extract information from textual resource and deals with operation such as retrieval, classification, clustering, data mining, natural language preprocessing and machine learning techniques together to classify different pattern. Text classification is the primary requirement of text retrieval systems, which retrieve texts in response to a user query .It, plays an important role in information extraction, summarization, and question answering. This paper is indented to deal with the text classification process using machine learning techniques. The machine learning methods for text classification such as Naïve Bayes, Support Vector Machine, K-Nearest Neighborhood, and Decision tree is used in different text dataset, and the performances of all these algorithms are compared and tabulated.

Keywords - **Text mining, Text Classification, Naïve Bayes, Support Vector Machine, K-Nearest Neighbor, Decision Tree, Precision, Recall, F- measure.**

I. INTRODUCTION

Text Classification [1] involves assigning a text document to a set of pre-defined classes automatically, using a machine learning technique. Text classification is a supervised learning technique that uses labeled training data to derive a classification system and then automatically classifies unlabelled text data using the derived classifier. The most data for text classification are collected from the web, through newsgroups, bulletin boards, and broadcast or printed news. Many classification methods have been developed with the aid of learning algorithms such as Naïve Bayesian, K-Nearest Neighbor (KNN), Support Vector Machine and Decision Tree. These classifiers are basic learning methods and adopt sets of rules.

II. TEXT CLASSIFICATION

Text classification [8] [9] is a fundamental task in document processing. The goal of text classification is to classify a set of documents into a fixed number of predefined categories/classes. . More formally, if d_i is a document of the entire set of documents *D* and $\{c_1, c_2, \ldots\ldots, c_n\}$ is the set of all the categories, then text classification assigns one category c_j to a document d_i..A document may belong to more than one class. The process of text classification systems can be separated into two main phases: information retrieval phase when numerical data is being extracted from the text and next is the main classification phase when an algorithm processes this data to make a decision on what category should the text belongs to. The stages of Text Classification[1] [4] [17] include the following process as shown in Fig. 1.

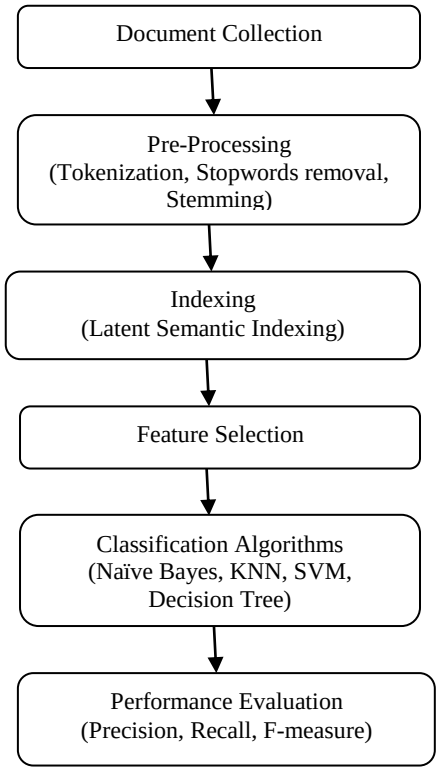

Figure 1: Text Classification Framework

The first step of Text Classification process is collecting the different types (format) of document like html, .pdf, .doc, web content etc. Next the pre-processing is used to present

the text documents into clear word format. The steps include tokenization, removing stop words and stemming. In indexing process, the Latent Semantic Indexing (LSI) technique is used which preserves the representative features for a document. After pre-processing and indexing the important step of text classification, is feature selection to construct vector space, which improves the scalability, efficiency and accuracy of a text classifier. The Text Classification algorithms are used to classify the documents in to predefined category based the class label. The last stage of Text classification is performance evaluation in which the evaluation is conducted experimentally. The performances of the classifiers are evaluated using many performance metrics such as Precision, Recall, F-measure, etc.

III. NAÏVE BAYES CLASSIFIER

Naïve Bayesian [2] [5] [13] is simple and efficient to implement as it assumes that all the words of the documents are independent to one another. The Naïve Bayes Classifier is the simplest probabilistic classifier used to classify the text documents. Naïve Bayes method is kind of module classifier under known priori probability and class conditional probability. The basic idea is to use the joint probabilities of words and categories to estimate the class of a given document. Given a document d_i , the probability with each class c_j is calculated as

$$P(c_j/d_i)=P(d_i/c_j).P(c_j)/P(d_i)$$

As $P(d_i)$ is the same for all class, then label(d_i) is the class (or label) of d_i , can be determined by

$$\text{label}(d_i)= \arg \text{Max}c_j\{ P(c_j /d_i)\}$$
$$= \arg \text{Max}\{P(c_j))/ P(d_i /c_j)P(c_j)\}$$

This technique Classify using probabilities and assuming independence among terms

$$P(C/X_i\ X_j\ X_k) = P(C)\ P(X_i/C)\ P(X_j/C)\ P(X_k/C)$$

IV. SUPPORT VECTOR CLASSIFIER

A Support Vector Classifier [10] [12] is a supervised classification algorithm that has been extensively and successfully used for text classification task. The SVM need both positive and negative training set which are uncommon for other classification methods. These positive and negative training set are needed for the SVM to seek for the decision surface that best separates the positive from the negative data in the n dimensional space, so called the hyper plane. The document representatives which are closest to the decision surface are called the support vector. SVM try to find an optimal hyper plane within the input space so as to correctly classify the binary (or multi-class) classification problem. For linearly separable space (i.e. for binary classification problem), the hyper plane is written as

$$w.x+b=0$$

Here x is an arbitrary object to be classified; the vector w and constant b are learned from a training set of linearly separable objects. In case of linearly separable data, SVM separates the positive and negative training examples with a maximum margin.

V. NEAREST NEIGHBOR CLASSIFIER

The K-Nearest Neighbor (KNN) algorithm [6] [14] is simple, valid and non-parameter method. KNN is also called instance-based learning or lazy learning. In this, each document is represented by nodes. For classification, distance between each labeled node (labeled document) and unlabeled node (unlabeled document) is calculated. And to decide whether the document (d_i) belongs to class, the similarity or dissimilarity to all documents in the training set is determined. The distance between two neighbors using Euclidean distance can be found using the given formula

$$\text{Dist}(X, Y) = \sqrt{\sum_{i=1}^{D}(X_i - Y_i)^2}$$

VI. DECISION TREE CLASSIFIER

Decision trees [11] [16] are trees that classify instances by sorting them based on feature values. Each node in a decision tree represents a feature in an instance to be classified, and each branch represents a value that the node can assume. Instances are classified starting at the root node and sorted based on their feature values. J48 tree builds the decision tree from labeled training data set using information gain and it examines the same that results from choosing an attribute for splitting the data. To make the decision the attribute with highest normalized information gain is used. Then the algorithm recurs on smaller subsets. The splitting procedure stops if all instances in a subset belong to the same class. Then the leaf node is created in a decision tree telling to choose that class.

VII. PERFORMANCE MEASURES

There are various methods to determine effectiveness or the performance of the algorithms. The metrics Precision, Recall, and F-measure are most often used.

Precision[2] [7][18] is determined as the conditional probability that a random document d is classified under ci, or what would be deemed the correct category.

$$Precision = \frac{TP}{TP+FP}$$

Recall is defined as the probability that, if a random document (dx) should be classified under category (ci), this decision is taken.

$$Recall = \frac{TP}{TP+FN}$$

where

True Positive (TP) - situation in text classification when the classifier correctly classifies a positive test case into the positive class;

True Negative (TN) – situation in text classification when the classifier correctly classifies a negative test case into the negative class;
False Positive (FP) – situation in text classification when the classifier incorrectly classifies a negative test case into the positive class;
False Negative (FN) – situation in text classification when the classifier incorrectly classifies a positive test case into the negative class;

Precision and recall are often combined in order to get a better picture of the performance of the classifier given as F-Measure [15]

$$\text{F} - \text{Measure} = \frac{2\times\text{Recall}\times\text{Precision}}{\text{Recall}+\text{Precision}}$$

VIII. EXPERIMENTS AND EVALUATION

The Performance metrics of the text classifiers such as Naïve Bayes (given as Naïve Bayes in Weka), Support Vector (SMO in Weka), K- Nearest Neighbor (IBk in Weka) and Decision Tree (J48 in Weka) are compared using different datasets. Two different text datasets namely breast-cancer dataset and hypothyroid datasets is used for the evaluation of the classifiers. The datasets are obtained from the Universal Client Identification (UCI) repository. The Weka tool is used for the evaluation of the classifiers on the different datasets and the metrics value obtained for the corresponding dataset on applying different classifiers is tabulated.

The breast-cancer dataset is taken first and the performance of the different classifiers is evaluated. The breast-cancer dataset contains 286 instances and the correctly classified instances and incorrectly classified instances with different classifiers are given in Table 1. It is shown that the Decision Tree classifier (J48) has higher value of 216 correctly classified instances compared to other classifiers. The percentage of correctly classified instances and incorrectly classified instance is shown in Fig. 2.

Table 1: Classified Instances of Breast-cancer Dataset

Terms	Correctly Classified Instances	Incorrectly Classified Instances
Naive Bayes	205	81
SMO	199	87
IBk	207	79
J48	216	70

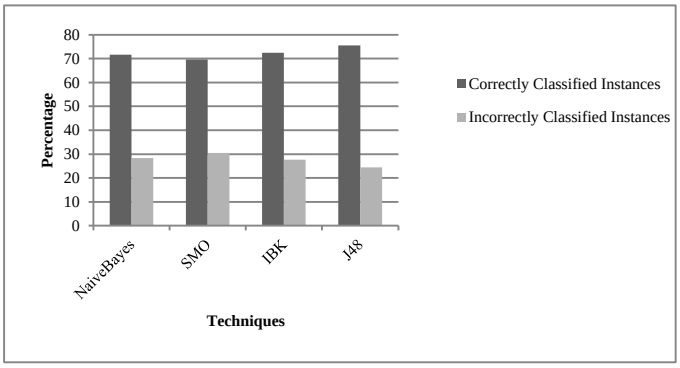

Figure 2: Percentage chart for Classified Instances of Breast-cancer dataset

The other performance metrics values such as TP, FP, Precision, Recall and F-measure for breast-cancer dataset using different classifiers is given in Table 2.It is observed that the Decision Tree Classifier (J48) has the high value measure rate, compared to other classifiers. The percentage value for the performance of breast-cancer dataset is shown in Fig. 3.

Table 2: Performance Metrics of Breast-Cancer dataset using different classifiers.

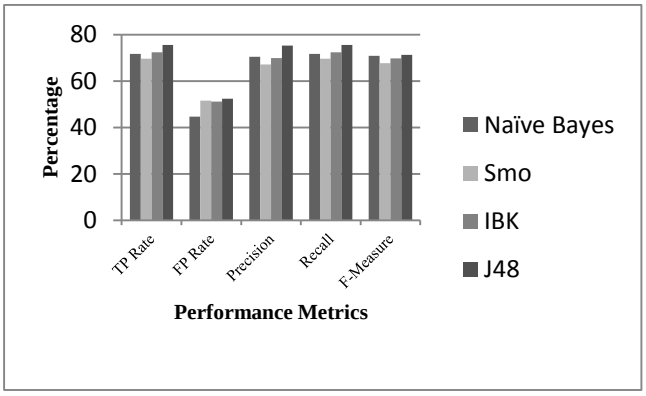

Figure 3: Percentage chart for performance of Breast - Cancer dataset.

The hypothyroid dataset is taken as the second dataset and the performance of the different classifiers is evaluated. The hypothyroid dataset contains 3772 instances and the correctly classified instances and incorrectly classified instances with different classifiers are given in Table 3. It is shown that the Decision Tree classifier (J48) has higher value of 3756 correctly classified instances

Terms	**TP Rate**	**FP Rate**	**Precision**	**Recall**	**F-Measure**
Naive Bayes	0.717	0.446	0.704	0.717	0.708
SMO	0.696	0.516	0.671	0.696	0.677
IBk	0.724	0.511	0.699	0.724	0.697
J48	0.755	0.524	0.752	0.755	0.713

compared to other classifiers. The percentage of correctly classified instances and incorrectly classified instance is shown in Fig. 4.

Table 3: Classified Instances of hypothyroid dataset.

Terms	Correctly Classified Instances	Incorrectly Classified Instances
Naive Bayes	3594	178
SMO	3531	241
IBk	3452	320
J48	3756	16

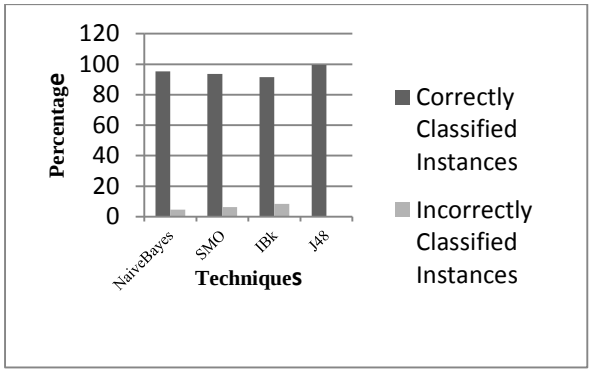

Figure 4: Percentage chart for Classified Instances of hypothyroid dataset.

The other performance metrics values such as TP, FP, Precision, Recall and F-measure for hypothyroid dataset using different classifiers is given in Table 4.It is observed that the Decision Tree Classifier (J48) has the high value of measure rate, compared to other classifiers. The percentage value for the performance of breast-cancer dataset is shown in Fig. 5.

Table 4: Performance Metrics of Hypothyroid dataset using different classifiers

Terms	TP Rate	FP Rate	Precision	Recall	F-Measure
Naive Bayes	0.953	0.448	0.946	0.953	0.945
SMO	0.936	0.755	0.888	0.936	0.91
IBk	0.915	0.573	0.908	0.915	0.911
J48	0.996	0.019	0.995	0.996	0.995

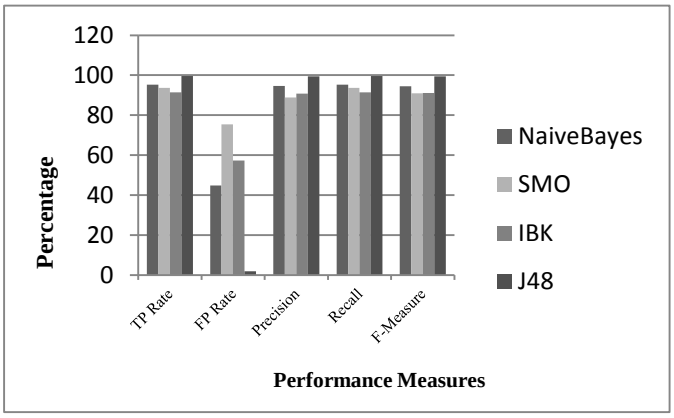

Figure 5: Percentage chart for Performance of Hypothyroid dataset.

XI. CONCLUSION

The machine learning techniques Naive Bayes, Support Vector Machine, K-Nearest Neighbor and Decision Tree for Text Classification is compared with each other on their performance. Based on the evaluation the Decision Tree classifier (J48) has the high precision, recall and F-measure value for both breast-cancer and hypothyroid datasets. It is observed that for specified classification method, classification performance of the classifiers based on different dataset ,the corpuses is different. From the above discussion it is inferred that no single representation scheme and classifier can be mentioned as a general model for any application. Different algorithms perform differently depending on the data collection.

REFERENCES

[1] C.Kanakslakshmi, Dr.R.Manicka chezian, "An Analysis on Text Mining and Text Classification Techniques", Proceedings of the National Conference on Information and Image Processing, Volume 1, Page .No 132-135,February 2015.

[2] G.Angulakshmi, Dr.R.Manicka Chezian, " Three Level Feature Extraction For Sentiment Classification", International Journal of Innovative Research in Computer and Communication Engineering, Volume 2, Issue 8, Page .No5501-5507, August 2014.

[3] Amerada Patra, Divakar Singh, "A Survey Report On Text Classification with Different Term Weighing Methods and Comparison between Classifications Algorithms", International Journal of Computer Applications, Volume 75, Issue7, Page .No14-18, August 2013.

[4] Aakanksha, Er.Dinesh kumar, "A Hybrid Approach For Text Classification Using Hmm, Svm And Genetic Algorithm", International Journal For Technological Research In Engineering , Volume 1, Issue 12, Page .No 1454-1457, August-2014.

[5] A. Saritha, N. NaveenKumar, "Effective Classification of Text", International Journal of Computer Trends

and Technology (IJCTT), Volume 11,Issue 1, Page .No 1-6, May 2014.

[6] J.Sreemathy, P.S.Balamurugan, "An Efficient Text Classification Using Knn And Naive Bayesian", International Journal on Computer Science and Engineering (IJCSE), Volume 4, Issue 3, Page .No 392-396, March 2012.

[7] Jafar Ababneh, Omar Almomani,, Wael Hadi, Nidhal Kamel Taha El-Omari, and Ali Al-Ibrahim, "Vector Space Models to Classify Arabic Text", International Journal of Computer Trends and Technology (IJCTT) ,Volume 7 , Issuer 4, Page .No 219-223, January 2014.

[8] Megha Gupta, Naveen Aggrawal, , "Classification Techniques Analysis", National Conference on Computational Instrumentation CSIO , Chandigarh, India, Page .No 128-131,March 2010.

[9] Mita K. Dalal, Mukesh A. Zaveri, "Automatic Text Classification: A Technical Review", International Journal of Computer Applications, Volume 28, Issue 2, Page .No 37-40, August 2011.

10] Nidhi,Vishal Gupta, "Recent Trends in Text Classification Techniques", International Journal of Computer Applications, Volume 35,Issue 6, Page .No 45-51, December 2011.

[11] Raj Kumar, Dr. Rajesh Verma, "Classification Algorithms for Data Mining:A Survey", International Journal of Innovations in Engineering and Technology (IJIET), Volume 1, Issue 2 , Page .No 7-14, August 2012.

[12] Rashedur M. Rahman, Farhana Afroz, "Comparision of Various Classifiaction Techniques Using Different Data Mining Tools for Diabetes Diagonosis", Journal of Software Engineering and Applications, Volume 6, Page .No 85-97, March 2013.

[13] S. Ramasundaram and S.P. Victor, "Algorithms for Text Categorization : A Comparative Study", World Applied Sciences Journal,Volume 22,Issue 9, Page .No 1232-1240, 2013.

[14] Sadegh Bafandeh Imandoust , Mohammad Bolandraftar, " Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background" ,International Journal of Engineering Research and Applications ,Volume 3, Issue 5, Page .No 605-610, October 2013.

[15] Shweta C. Dharmadhikari,Maya Ingle,Parag Kulkarni, "A Comparative Analysis of Supervised Multi-label Text Classification Methods", International Journal of Engineering Research and Applications (IJERA) Vol. 1, Issue 4, Page .No 1952-1961, March 2012.

[16] Trilok Chand Sharma1, Manoj Jain, "WEKA Approach for Comparative Study of Classification Algorithm", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 4, Page .No 1925-1931, April 2013.

[17] Vaibhav, C.Gandhi, Jignesh A.Prajapati, "Review on Comparison between Text Classification Algorithms", International Journal of Emerging Trends & Technology in Computer Science, Volume 1, Issue 3, Page .No 75-78, October 2012.

[18] Zakaria Elberrichi, Abdelattif Rahmoun, Mohamed Amine Bentaalah , "Using Word Net for Text Categorization" , International Arab Journal of Information Technology, Volume 5, Issue 1, Page .No16-24, January 2008.

BIOGRAPHIES

Dr. R.Manickachezian received his M.Sc., degree in Applied Science from P.S.G College of Technology, Coimbatore, India in 1987. He completed his M.S. degree in Software Systems from Birla Institute of Technology and Science, Pilani, Rajasthan, India and Ph D degree in Computer Science from School of Computer Science and Engineering, Bharathiar University, Coimbatore, India. He served as a Faculty of Maths and Computer Applications at P.S.G College of Technology, Coimbatore from 1987 to 1989. Presently, he has been working as an Associate Professor of Computer Science in N G M College (Autonomous), Pollachi under Bharathiar University, Coimbatore, India since 1989. He has published thirty papers in international/national journal and conferences. He is a recipient of many awards like Desha Mithra Award and Best Paper Award. His research focuses on Network Databases, Data Mining, Distributed Computing, Data Compression, Mobile Computing, Real Time Systems and Bio-Informatics.

C.Kanakalakshmi is a Research Scholar in Department of Computer Science, Nallamuthu Gounder Mahalingam College, Pollachi. She received her Master of Computer Applications (M.C.A) in 2011 from Nallamuthu Gounder Mahalingam College, Pollachi under Bharathiar University, Coimbatore. She has presented papers in International/National conferences and attended Workshop, Seminars. Her research focuses on Data Mining.