DOI: http://dx.doi.org/10.26483/ijarcs.v8i8.4699

ISSN No. 0976-5697

Volume 8, No. 8, September-October 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

EMPIRICAL EVALUATION OF MACHINE LEARNING ALGORITHMS FOR AUTOMATIC DOCUMENT CLASSIFICATION

P.V.Arivoli
Research Scholar, Department of Computer Science,
A.V.V.M. Sri Pushpam College, Poondi,
Thanjavur, India

T.Chakravarthy
Associate Professor, Department of Computer Science,
A.V.V.M. Sri Pushpam College, Poondi,
Thanjavur, India

G. Kumaravelan
Assistant Professor, Department of Computer Science,
Pondicherry University, Karaikal Campus,
Karaikal, India

Abstract: Automatic document classification process is the important area of research in the field of Text Mining(TM). Text mining is the process of discovering the interesting pattern or knowledge from huge amount of data. The document classification process used in many domains. Here, to take the classification process is apply SMS spam classification. The benchmarked dataset is used and the same data set is processed in various ML algorithms of Naïve Bayes, Support Vector Machine, Decision Tree and Logistic Regression model. In this paper evaluates the results of various machine learning algorithms for automatic document classification in SMS spam classification.

Keywords: Text Mining, Machine Learning, Document Classification and Information Retrieval.

I. INTRODUCTION

Automatic text document classification is the one among a prime functionality in the field of Text Mining area due to the exponential growth of unstructured data in the current digital era. The primary objective of classification functionality is to assign each document a predefined label automatically based on its contents. It is widely used in knowledge extraction and knowledge representation in text data sets. The well known applications which employs document classification functionalities are email categorization, spam filtering, directory maintenance, mail routing, news monitoring and narrow casting, etc.

In general, the text document classification process includes the two major phases namely, *document representation* and *classification process*. The document representation process is divided into two steps. They are feature extraction and feature selection. The feature extraction involves various preprocessing activities to reduce the document complexity and make the classification process in easier manner. Usually, the preprocessing process incorporates the stop word removal, stemming of words, punctuation removal and finally tokenization process. The feature extraction process includes the calculation of *Term Frequency* (TF) and *Inverse Document Frequency* (IDF) from the tokenized documents. Finally, all the documents are normalized to unit length. The second phase of document classification is the application of machine learning algorithms. Many machine learning algorithms are available like supervised, semi supervised and unsupervised learning algorithms. This paper focuses on supervised machine learning algorithms like Naive Bayes (NB), Support Vector Machine (SVM), K- Nearest Neighbor (K-NN), Decision Tree (DT) and Logistic Regression (LR) in automating the text document classification. The rest of the paper is organized as follows. Section 2 discuses about various machine learning algorithms used for classification process. This is followed in Section 3 by some experiments on SMS spam classification task. Finally, Section 4 concludes the paper.

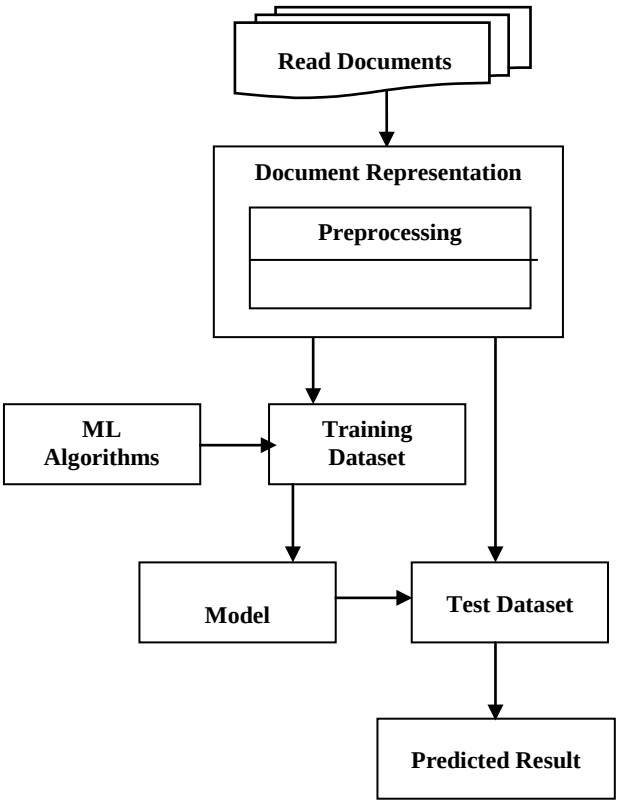


Figure 1. Text Document Classification Process Using ML Algorithms

II. DIFFERENT TYPES OF APPROACHES

Ethem Alpaydin defines Machine Learning (ML) is a paradigm which "optimize a performance criterion using example data or past experience" [7]. Machine learning is the intersection of computer science, engineering, and statistics and often appears in other disciplines. Machine learning uses statistics to solve many classification and clustering problems. The ML algorithms are classified in three categories. They are supervised, unsupervised and semi supervised. Now we discuss about few machine learning algorithms, like, Naïve Bayes (NB), Support Vector Machine (SVM), K- Nearest Neighbor (K-NN), Decision Tree (DT) and Logistic Regression (LR).

A. Naïve Bayes Classification:

The Naive Bayes (NB) classifier is classical and probabilistic classifier. It is a supervised learning technique of ML. It support only on numeric and textual data [2], [6], [13],[17], [20]. NB focuses on text document classification process and many application areas like detection of spam email, sorting personal email, classification of documents, language recognition and recognition of sentiment analysis. The merits of Naive Bayes are simple, fast and very effective; to eliminate the noisy and missing data values. Easy to capture the probability estimation of a prediction. Some demerits are fall on an often supposition of equally important and independent features; No ideal datasets with many features of numeric and less reliable than the predicted classes of estimated probabilities.

B. Support Vector Machines:

The SVM which works on the basis of statistical based method and also a supervised learning technique of ML. It is mainly used to solve the problems of regression and categorization [3], [18], [19],[23]. It's using a sigmoid kernel function is alike two-layer perceptron. A given class members of n-dimensional vectors and it is used to discriminate positive and negative, the training set supports both positive and negative. Computational learning theory that performs the structural risk minimization. SVM advantages are it can be used for classification or prediction of numeric problems. Not overly influenced by noisy data and not very prone to over fitting. It is easy to use than artificial neural networks, specifically due to the existence of many well-supported SVM models. Some disadvantages of SVM are the training is very slow, in case of input dataset has a huge feature. The result represented in a complex black box model. Find the best model, used to various combinations of kernels.

C. K-Nearest Neighbour:

The k-NN is supervised learning algorithm and also a non-parametric regression algorithm for text categorization [1],[4],[5],[8],[15],[21]. It is a first typical approach, classifies new cases based on a similarity measure, i.e. by using distance functions. By using some similarity measure such as Euclidean distance measure, etc., the distance is calculated by the Euclidean formula, as in Eq. (1)

$$\text{Dist}(x,y) = \sqrt{(xi - yi)2} \qquad (1)$$

The merits of KNN are Simple and effective, makes no assumptions about the underlying data distribution and performs well in the training phase. Its demerits is in classification phase it works very slow and requires some special cases while handling some missing data in the training phase.

D. Decision Trees:

A decision tree model to support the decisions and their possible outcome, including fortuity results, resource costs, and usefulness. It's like a tree structure and hierarchical structure with the acyclic directed graphs as shown in figure 2; The starting node is always a root node and the root node connects directly to the next level nodes. Final nodes (leafs) represents the categories of document, the tree leaf nodes hold examine the categorized documents should and travel all the nodes in order [10], [12], [14], [16]. Branches link nodes of adjacent levels, and then the testing process is executes on the selected document attributes. The test results are connected to branches proceeds to specific nodes of the bottom level. It can be focus the connections in specific nodes, reflect as an influence diagrams.

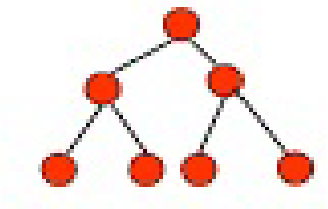


Figure 2. Decision Tree

The strength of decision tress is to receives all-intent classifier that does well on many problems, the automatic process skill is high, it accepts the numeric and ostensible features, to avoids the missing data trivial features. It supports on both small and huge datasets. Weaknesses of decision tree are models are splits on features in huge number of levels often biased. In large tree is easy to over fit or under fit the model. It can be critical to interpret and decisions they make may seem counterintuitive.

E. Logistic Regression:

Logistic regression is a powerful statistical model. In this model produces a binomial result of one or more descriptive variables. It calculates the relationship between the classification dependent variable and self-determining variables. Logistic function is used to estimating probabilities, which is the consolidate logistic distribution. [9],[11],[22].

III. EXPERIMENTAL SETUP:

A. Data set:

As worldwide use of mobile phones has grown, a new boulevard for electronic scrap mail has opened for scandalous vendors. These publicists exploit Short Message Service (SMS) text messages to target impending customers with unsolicited publicizing known as SMS junk. This type of junk is mostly troublesome because, several cellular phone users pay a fee per SMS acknowledged. Thus, classification process becomes evitable that could filter SMS junk would offer a valuable tool for cellular phone beneficiaries.

The benchmarked SMS spam collection is downloaded from http://www.dt.fee.unicamp.br/~tiago/smsspamcollection/

The SMS spam dataset has 5559 SMS messages with 2 features namely type and text. The SMS type has been coded

as either ham or spam. The text element stores the full raw SMS text. In the experiment setup 75% of the SMS spam dataset is fixed as the training dataset and remaining as the test dataset.

B. Experimental Results:

The outcomes of the experiments are visualized in the form of confusion matrix which shows the relationship between the positive and negative predictions of the class labels according to the given experimental design setup with one of the following grouping.

- True Positive (TP): Rightly classified as the class of relevance
- True Negative (TN): Rightly classified as not the class of relevance
- False Positive (FP): Wrongly classified as the class of relevance
- False Negative (FN): Wrongly classified as not the class of relevance

Figure 3 depicts the above said properties for the SMS spam classification task. Specifically, a confusion matrix is also called error matrix, is distinct table layout that allows visualization of the performance of the applied machine learning algorithms with benchmark dataset, confusion matrix. Various important statistical measures like accuracy, error, precision agreement, precision error, kappa statistics, sensitivity, specificity, precision, recall and F-measure are calculated from the resultant confusion matrix. According to our experimental procedure, the positive class is spam, which is our point of interest of the prediction.

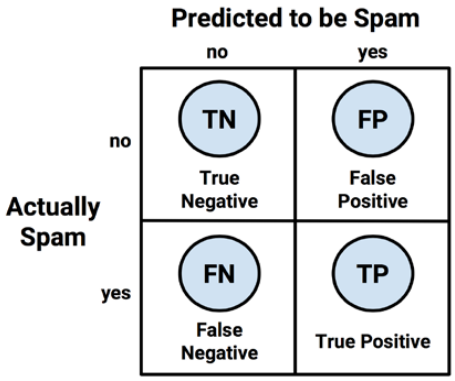


Figure 3. SMS spam classification

a) ***Accuracy rate****= TN+TP/ N , where N is the total of the classified items.*

b) ***Error rate*** *= FN+FP / N , where N is the total of the classified items.*

c) ***Kappa statistics*** *= Pr(a) - Pr(e) /1-Pr(e), where Pr(a) is the proportion of the actual agreement and Pr(e) refers to the expected agreement between the classifier and the true values.*

d) ***Sensitivity*** *= TP / TP+ FN*

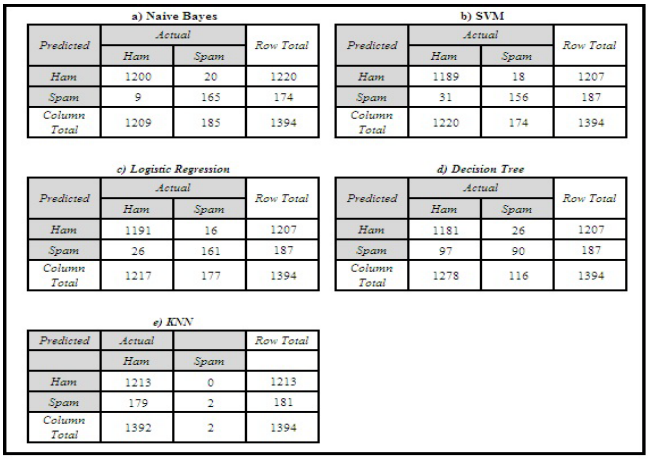
e) ***Specifisity*** *= TN / TN + FP*

f) ***Precision rate*** *= TP/TP+FP , the ratio of correctly classified items to all items classified to that class.*

g) ***Recall rate*** *= TP/TP+FN , the ratio of correctly classified items to all items of that class.*

h) ***F – measure*** *= 2 X Precision X Recall / Recall + Precision (or) 2 X TP / 2 X TP +FP + FN*

Figure 4 shows the screen shot of the respective confusion matrices of the applied machine learning algorithms on the benchmark dataset. From the respective confusion matrix the following statistical measures are empirically verified.



a) Naive Bayes

Predicted	Actual		Row Total
	Ham	Spam	
Ham	1200	20	1220
Spam	9	165	174
Column Total	1209	185	1394

b) SVM

Predicted	Actual		Row Total
	Ham	Spam	
Ham	1189	18	1207
Spam	31	156	187
Column Total	1220	174	1394

c) Logistic Regression

Predicted	Actual		Row Total
	Ham	Spam	
Ham	1191	16	1207
Spam	26	161	187
Column Total	1217	177	1394

d) Decision Tree

Predicted	Actual		Row Total
	Ham	Spam	
Ham	1181	26	1207
Spam	97	90	187
Column Total	1278	116	1394

e) KNN

Predicted	Actual		Row Total
	Ham	Spam	
Ham	1213	0	1213
Spam	179	2	181
Column Total	1392	2	1394

Figure 4. Confusion Matrices of of the applied machine learning algorithms on the SMS spam dataset.

Table I. Performance Summary of Machine Learning Algorithms for Document Classification

Model Name	*Accuracy Rate*	*Error Rate*	*Precision Agreement*	*Precision Error*	*Kappa Statistics*	*Sensitivity*	*Specificity*	*Precision*	*Recall*	*F-Measure*
Naïve Bayes	0.97919	0.02080	0.97919	0.77559	0.90729	0.94827	0.98360	0.89189	0.94827	0.91922
SVM	0.96198	0.03802	0.96198	0.77662	0.82979	0.81283	0.98508	0.89411	0.81283	0.85154
Logistic Regression	0.95767	0.04232	0.95767	0.79026	0.79819	0.72727	0.99337	0.94444	0.72727	0.82175
Decision Tree	0.90674	0.09325	0.90674	0.76034	0.61086	0.68984	0.94034	0.64179	0.68984	0.66494
KNN	0.871593	0.128407	0.871593	0.864127	0.054947	0.032432	1	1	0.032432	0.062827

Table I shows the performance results of SMS spam classification task using various machine learning algorithms, like of Naive Bayes, support vector machine, decision tree, k-nearest neighbor and logistic regression. The accuracy rate of decision tree is 90.67%. Logistic regression model accuracy rate is 95.76%, it increases by 5.09% from the existing model of decision tree. The SVM produces the 96.19 % of accuracy it increases the performance by 0.43% of logistic regression model and 5.52% of decision tree. The accuracy rate of KNN is 87.12%. In overall comparisons Naïve Bayes model outperforms the benchmarked algorithms and gives the accuracy of 97.92%.

IV. CONCLUSION

This paper investigated the state-of-the-art machine learning algorithms in text document classification. A comparison between them was also conducted in correspondence to the benchmark SMS spam dataset to find the SMS messages are either ham or spam using the well-established statistical measures like accuracy, kappa statistics, sensitivity, specificity, precision, recall and F- measure. At the nutshell the statistical classifier Naïve Bayes algorithms shows better performance in all categories.

V. REFERENCES:

[1] A.Kousar Nikhath, K.Subrahmanyam, R.Vasavi, "Building a K-Nearest Neighbor Classifier for Text Categorization", International Journal of Computer Science and Information Technologies, Vol. 7 No.1, pp. 254-256, 2016.

[2] Andrew McCallum and Kamal Nigam, “A Comparison of Event Models for Naïve Bayes Text Classification”, AAAI-98 workshop on learning for text categorization, Vol. 752, 1998.

[3] Arivoli. P.V., Chakravarthy. T, “Document Classificaiton Using Machine Learning Algorithms – A Review”, International Journal of Scientific Engineering and Research, Vol 5, Issue 2, pp 48 -55, February 2017.

[4] Bang, S. L., Yang, J. D., and Yang, H. J. , “Hierarchical document categorization with k-NN and concept-based thesauri, Elsevier, Information Processing and Management”, Vol. 42 No.2, pp. 397–406, 2006.

[5] Duoqian Miao , Qiguo Duan, Hongyun Zhang and Na Jiao, “Rough set based hybrid algorithm for text classification”, Elsevier, Expert Systems with Applications, Vol. 36, Issue 5, pp. 9168–9174, July 2009.

[6] El Kourdi, M., Bensaid, A., & Rachidi, T. E. , “Automatic Arabic document categorization based on the Naïve Bayes algorithm” In Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, Association for Computational Linguistics, pp. 51-58, August 2004.

[7] Ethem Alpaydin, “Introduction to Machine Learning (Adaptive Computation and Machine Learning)”, The MIT Press, 2004.

[8] Eui-Hong (Sam) Han, George Karypis and Vipin Kumar, “Text Categorization Using Weighted Adjusted k-Nearest Neighbor Classification”, Pacific-asia conference on knowledge discovery and datamining. Springer, Berlin, Heidelberg, pp.53-65,2001.

[9] Genkin, A., Lewis, D. D., & Madigan, D. “Large-scale Bayesian logistic regression for text categorization. Technometrics”, American Statistical Association and the American Society for Quality TECHNOMETRICS, Vol. 49, No. 3,pp. 291-304, 2007. DOI:10.1198/004017007000000245.

[10] Hwee-Tou Ng, Wei-Boon Goh and Kok-Leong Low , “Feature Selection, Perceptron Learning, and a Usability Case Study for Text Categorization, In Proceedings of the 20th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, pp.67-73. 1997.

[11] Ismail Hmeidi, Mahmoud Al-Ayyoub, Nawaf A. Abdulla, Abdalrahman A. Almodawar, Raddad Abooraig, Nizar A. Mahyoub, "Automatic Arabic text categorization: A comprehensive comparative study", Journal of Information Science,

[12] Kim. J, Lee. B, Shaw. M, Chang. H and Nelson. W, “Application of Decision -Tree Induction Techniques to Personalized Advertisements on Internet Storefronts”, International Journal of Electronic Commerce Vol .5 No.3, pp.45-62, 2001.

[13] Moromi Gogoi and Shikhar Kumar Sarma, "Document Classification of Assamese Text Using Naïve Bayes Approach", International Journal of Computer Trends and Technology (IJCTT), Vol. 30, No. 4, December 2015.

[14] Russell Greiner and Jonathan Schaffer, “AIxploratorium – Decision Trees”, Department of Computing Science, University of Alberta, Edmonton, ABT6G2H1, Canada.2001. URL :http://www.cs.ualberta.ca/ ~aixplore/ learning/ DecisionTrees

[15] S.G. Lade and Nikhil Vyawahare, "Document Classification Using KNN on GPU", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Vol. 3 Issue 8, August 2014.

[16] Said Bahassine, Abdellah Madani, Mohamed Kissi, "Arabic Text Classification Using New Stemmer For Feature Selection And Decision Trees", Journal of Engineering Science and Technology,Vol. 12, No. 6, pp. 1475-1487, 2017.

[17] Sang-Bum Kim, Kyoung-Soo Han, Hae-Chang Rim, and Sung Hyon Myaeng, "Some Effective Techniques for Naïve Bayes Text Classification", IEEE Transactions On Knowledge And Data Engineering, Vol. 18, No. 11, November 2006.

[18] Saurav Sahay, “Support Vector Machines and Document Classification”,URL:http://www.static.cc.gatech.edu/~ss ahay/sauravsahay7001-2.pdf . 2011.

[19] Thorsten Joachims, “Text Categorization with Support Vector Machines: Learning with Many Relevant Features” ECML -98, 10th European Conference on Machine Learning, pp. 137-142, 1998.

[20] Vishwanath Bijalwan, Pinki Kumari, Jordan Pascual and Vijay Bhaskar Semwal, "Machine learning approach for text and document mining", https://arxiv.org/ftp/arxiv/papers/1406/1406.1580.pdf, 2014.

[21] Vishwanath Bijalwan, Vinay Kumar, Pinki Kumari and Jordan Pascual, "KNN based Machine Learning Approach for Text and Document Mining", International Journal of Database Theory and Application, Vol.7, No.1, pp.61-70, 2014.

[22] Vladimir N. Vapnik, “The Nature of Statistical Learning Theory” , Springer science & business media, 2013.

[23] Yiming Yang and Xin Liu, "A re-examination of text categorization methods", In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 42-49, 2009. doi>10.1145/312624.312647.