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 Naïve 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 discusses 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.
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 Naïve 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 Naïve 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)

\[
\text{Dist}(x,y) = \sqrt{(x1 - y1)^2}
\]

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.

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.

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) Specificity = 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.

![Figure 3. SMS spam classification](image)

![Figure 4. Confusion Matrices of of the applied machine learning algorithms on the SMS spam dataset.](image)

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Accuracy Rate</th>
<th>Error Rate</th>
<th>Precision Agreement</th>
<th>Precision Error</th>
<th>Kappa Statistics</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>0.97919</td>
<td>0.02080</td>
<td>0.97919</td>
<td>0.77559</td>
<td>0.90729</td>
<td>0.94827</td>
<td>0.98360</td>
<td>0.89189</td>
<td>0.94827</td>
<td>0.91922</td>
</tr>
<tr>
<td>SVM</td>
<td>0.96198</td>
<td>0.03802</td>
<td>0.96198</td>
<td>0.77662</td>
<td>0.82979</td>
<td>0.81283</td>
<td>0.98508</td>
<td>0.89411</td>
<td>0.81283</td>
<td>0.85154</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.95767</td>
<td>0.04232</td>
<td>0.95767</td>
<td>0.79026</td>
<td>0.79819</td>
<td>0.72727</td>
<td>0.99337</td>
<td>0.72727</td>
<td>0.72727</td>
<td>0.82175</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.90674</td>
<td>0.09325</td>
<td>0.90674</td>
<td>0.76034</td>
<td>0.61086</td>
<td>0.68984</td>
<td>0.94034</td>
<td>0.64179</td>
<td>0.68984</td>
<td>0.66494</td>
</tr>
<tr>
<td>KNN</td>
<td>0.871593</td>
<td>0.128407</td>
<td>0.871593</td>
<td>0.864127</td>
<td>0.054947</td>
<td>0.032432</td>
<td>1</td>
<td>1</td>
<td>0.032432</td>
<td>0.062827</td>
</tr>
</tbody>
</table>
Table 1 shows the performance results of SMS spam classification task using various machine learning algorithms, like of Naïve 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 is 95.76%, it increases by 5.09% from the existing model of decision tree. The SVM produces the 96.19 % of accuracy it 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:


