International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

Feasible Performance Comparison of E-mail Spam Classification Based on Machine Learning Techniques

Mitu Pal[1], Bristi Rani Roy[2]

[1]*mitu151350@gmail.com, Lecturer, Dept. of CSE, Haji Abul Hossain Institute of Technology, Bangladesh*
[2]*bristiranyroy@gmail.com, Lecturer, Dept. of CSE, Bangladesh Army University of Engineering & Technology, Bangladesh*

Abstract - *Worldwide email is a common and fast communicating way and relatively low sending cost for message transfer protocol. But sometimes without filtering mail box are fill-up with unsolicited bulk email and junk email that is known as spam email. Many financial transaction and electronic business contribute or promote their business through email, which is very annoying to users. The use of spam email is rapidly increasing day after day. For that reason, filtering is essential and popular one to stop spam email. ML approaches are given more successful rate to filtering the spam email. In our paper, we give an overview some of ml classification algorithms as K-Nearest Neighbor (KNN), Naive Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), Multilayer perception (MLP) are used for learning the features of spam emails. By using the confusion matrix on 10-fold cross-validation in this paper to compare the performance of those six ML classifiers based on accuracy, recall & precision. The main goal of this article is to determine the better spam classification techniques for spam detection.*

***Key Words*: K-Nearest Neighbors, Naïve Bayes, Support Vector Machine, Logistic Regression, Random Forest, Multilayer perception, Accuracy, Precision, Recall, ROC Curve Analysis.**

1. INTRODUCTION

Now a day's internet has become an integral parts of our daily life. It is growing lavishly day by day. We exchange information through internet using different tools, due to it takes less time and efficient also low cost. E-mail is one of the mostly used tools for information exchange. Email provides some advantages over other method such as, data security during information exchange, negligible time delay, low cost etc. But there is some issues that spoil the pleasure of using email efficiently. And what can be a great example of it than spam. Unsought bulk of junk email is called spam email. On the internet it is a massive problem. In recent statistics, 40% of all emails are spam which about 15.4 billion email per day and that cost internet users about $355 million per year [1]. Spam email is very cheap to send so that, a large number of spam email is sent to the users. When large number of spam email is received by users then it is very hard to detect spam or ham email and also it takes time to delete during in this time period it may crash the server. It causes many problem for users such as waste of time, storage, computational power, money laundering etc. Spam filtering is one of the effective way to detect spam email. But spammers now a days use tricky method to pass filtering successfully. However knowledge engineering and machine learning is still effective than filtering to detect spam email. Machine learning approach does not require specifying any rules that's why Machine learning approach is more efficient than knowledge engineering approach [2]. The mail goal of the article is to detect spam email with high accuracy using different Machine Learning (ML) classification approaches. Rest of the article is indexed as follows: in section 2 we discuss the summary of related paper. Section 3 explain the dataset. We discuss about different ML classification techniques in section 4. In section 5 we analysis the experimental result. We show the comparison of different ML techniques in section 6. In section 7 we enclose the paper with conclusion.

2. RELATED RESEARCH WORK

Many research has been done for spam email detection using ML techniques or other techniques. Here we try to summarize some related work for spam classification.

In [1] authors used different ML classification technique for spam classification task. They used SVM, NB, KNN, AIS, NN, RS algorithm for spam detection.

In [3] authors proposed a model using the SVM for classification task. Here they analyze sender behavior and give a trust value based on this trust value they classify spam email. They also show that SVM classifier is effective than Random Forest.

In [4] authors have used neural network approach for spam email classification task. Though from the result we can show that ANN achieved good accuracy and it is good for spam classification but it is not efficient as a spam filtering tool to be used ANN alone.

In [5] authors have used Neural Network, SVM classifier, Naïve Bayesian Classifier, and J48 four different classifier for spam classification task. They applied these approaches based on different feature size and also for different data size.

3. ILLUSTRATION OF DATASET

From UCI Machine Learning Repository we can gather all types of datasets for machine learning techniques. Spam dataset also collected from UCI that is consists of 4601 email messages and there are 58 attributes for each instance. The 57 criterions are as discussed classification section, they are meant to determine if a message is spam or non-spam, the last feature contains a binary value (1 for spam email and 0 for non-spam email). In 57 attributes, it represent the frequency of a given word or character in the email. There is no missing value in spam dataset.

i. W_f_w: 48 attributes telling the frequency of word **w**, the percentage of words in the email, i.e. (number of times the **w** appears / total number of words in e-mail)*100.

ii. C_f_c: 6 attributes describing the frequency of a character **c**, percentage of characters in the email, i.e. (number of **c** occurrences / total characters in e-mail)*100.

iii. C_f_cap: 3 attributes describing the longest length, total numbers of capital letters and average length.

iv. Class: last 1 attribute denotes whether the e-mail was considered spam (1) or not (0).

4. MACHINE LEARNING METHODS

Machine learning techniques are divided into two categories: supervised and unsupervised classification methods are used to train the machine and show correct result in supervised learning. In unsupervised learning, datasets are not pre-determined. ML methods can employ statistics, probabilities, Boolean logic, and unconventional optimization to classify patterns or to build prediction models [6]. We used several machine learning techniques in our study. All these approaches have been described in this section.

4.1 K-Nearest Neighbors

K-Nearest Neighbors is a simple non-parametric method for classifying cases based on other similarity cases. When new data points get classified it take an individual class in case of classification. It is a supervised learning algorithm.

To predict the outcome of a new instance, we use the Euclidean distances to evaluate the distance between the instance and all the points in the training set. Euclidean distance is given in equation (1). There are some other distance calculation metrics as well named Manhattan distance, city block distance etc. In this correspondence Euclidian distance is used,

$$D = \sqrt{\sum_{i=1}^{n}(xi - yi)^2}$$

Model as,

i. Input the dataset.

ii. Split the dataset into training and testing set.

iii. Fit the model of training data.

iv. Calculate the score of testing data.

4.2 Naïve Bayes

Naïve Bayes is a classification technique based on Bayes' Theorem with an assumption of independent among predictors. The building process of Navies Bayes model is very simple and especially useful for large data set and extremely sophisticated classification method.

Bayes theorem calculates the posterior probability, the equation is given below:

$$p\left(\frac{i}{y}\right) = \frac{p(y/_i)p(i)}{p(y)}$$

i. P (i|y) is the posterior probability of class (i, target) given predictor(y, attribute) which represents the degree to which we believe a given model accurately describes the situation given the available data and all of our prior information.

ii. P(i) is the prior probability of class which describes the degree to which we believe the model accurately describes reality based on all of our prior information

iii. P (y|i) is the likelihood which describes how well the model predicts the data

iv. P(y) is the probability of predictor

4.3 Support Vector Machine

SVM algorithm is one kind of machine learning algorithm which can be used for both regression and classification [6]. It is done through putting a line in the Cartesian plane to separate the types of data. This line is called the hyper plane.

A hyper plane is a line that intersects the input variable space. In SVM, a hyper plane is selected to best separate the points in the input variable space by their class.

4.4 Logistic Regression

Logistic regression is the appropriate regression and a statistical procedure for exploring a dataset to produce measurements to a discrete set of groups [7] in which there are one or more independent variable that determine an effect and the dependent variable is dichotomous i.e. it only includes data codes as 1 (True) or 0 (False). The aim of logistic regression is to find the best suitable model to interpret the relationship among the dependent and a set of independent variable.

4.5 Random Forest

Random forest builds multiple decision trees and merges them together to get a more accurate prediction [7]. Random forest can equally use for solved both classification and regression problem. Sometimes over-fitting may generate the results worse, but for Random Forest there are enough trees in the forest, so there is no need to change for spam classification if we do so then it will over-fit the model. Random forest can handle lost or missing values perfectly and if a new data is certainly enter into the dataset, it may not affect the whole algorithm. Random Forest classifier can be modeled for categorical values.

4.6 Multilayer Perception

Multilayer perceptron is an Artificial Neural Network (ANN) that is compose with more than one perceptron to solve depth problem. MLP consists of three layers. The first layer (input layer) receive the signal and send the output to the second layer (hidden layer), it also send desire output to the last layer (output layer). The output layer makes a decision or prediction about the input layer, also compared with the target output. And when a signal is error, back propagation is also used, because MLP is also known as Back Propagation Neural Network (BPNN) and use a supervised learning technique. In BPNN there are multiple layer of neurons (input, hidden and output layers), each neuron has connected with weights correspondingly during training and original result is balance with target value to complete the classification.

5. RESULT ANALYSIS

The whole dataset was split into two sets: one is a training set (80%) and another one is the testing set (20%). To train the model, we applied different Machine Learning techniques in spam dataset. By using 10 cross-validation the overall accuracy can be measured of this model. Accuracy, Precision and Recall are used for measuring the ML classifiers. In all this illusion, tp, tn and fp, fn represent true positive, true negative and false positive, false negative respectively, which have been explained below.

5.1 Precision

Proposition of correct positive classification (true positives) from cases that are predicted as positive. Precision is also commonly known as confidence. It is shown here that how precision handle positive observation to classify. High precision related to the low false positive rate.

$$\text{Precision} = \frac{tp}{tp + tn}$$

Table -1: Precision rate corresponding six ML techniques

Algorithms	Spam (1)	Non-spam (0)	Avg/total
KNN	0.79	0.86	0.83
NB	0.71	0.97	0.86
SVM	0.92	0.93	0.93
LR	0.91	0.92	0.92
RF	0.95	0.95	0.95
MLP	0.89	0.93	0.91

5.2 Recall

Recall is the ratio of correct predicted positive observation to the all observations in actual class. It is also commonly known as sensitivity.

$$Reall = \frac{tp}{tp+fn}$$

Table -2: Recall rate corresponding six ML techniques

Algorithms	Spam (1)	Non-spam (0)	Avg/total
KNN	0.80	0.86	0.83
NB	0.73	0.97	0.83
SVM	0.89	0.95	0.93
LR	0.88	0.94	0.92
RF	0.92	0.97	0.95
MLP	0.88	0.93	0.91

5.3 Accuracy

Accuracy is the most intuitive performance measure for evaluating classification models. Accuracy is the fraction of predictions for our model getting right. It is simply a rate of accurately predicted observation to the total observation. Accuracy can also be calculated by using positives and negatives terms for binary classification.

$$Accuracy = \frac{tp+tn}{tp+tn+fp+fn}$$

Best model can be getting when we have high accuracy. Accuracy is a great measure but only when the values of false positive and false negatives are almost same.

Table -3: Overall Accuracy rate for corresponding six ML techniques.

Algorithms	Accuracy (%)
KNN	83%
NB	83%
SVM	93%
LR	92%
RF	95%
MLP	91%

6. PERFORMANCE EVALUATION

Performance comparison is one of the most important tasks to determine the effective approach for any classification or other task. In this section, we will compare the performance of different ML classification approaches to determine which approach is better for spam classification. For doing that we will use the accuracy graph and Roc curve.

From that accuracy graph, we can show RF achieves better accuracy than other approaches. Though SVM and MLP also perform pretty well but not as well as RF. And NB, KNN doesn't perform well compared to RF, NB, and MLP. We can say that for large dataset RF perform well. And also for spam classification, RF is more effective compare to other techniques or approaches.

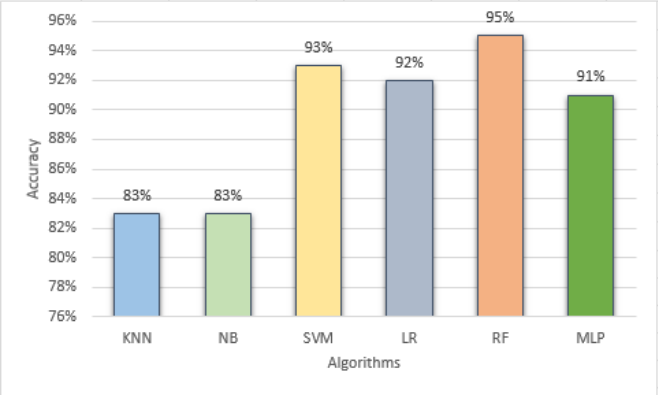

Figure-1: Accuracy comparison using six ML techniques

ROC Curve Analysis

A receiver operating curve or ROC curve is used for performance comparison of different approaches. Here mainly based on AUC value performance is compared. The higher the AUC value better the techniques perform well. Here we also show the performance comparison based on the ROC curve.

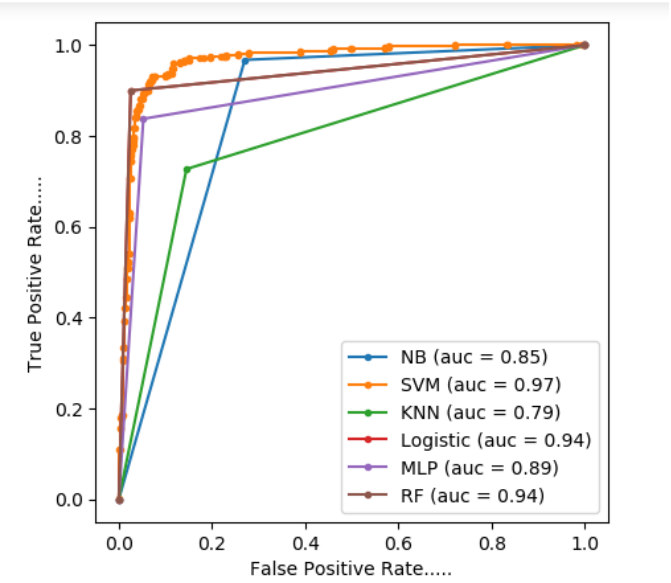

Figure-2: Evaluating the ROC curve using six ML techniques based on accuracy

From this ROC curve, we can show that SVM has achieved the highest AUC value compared to other approaches. After this LR and RF have achieved pretty well accuracy compared to MLP, NB, and KNN. However, NB and KNN achieve not satisfying accuracy compared to other techniques. In this case, SVM and RF perform well for spam detection with satisfying accuracy and AUC value.

7. CONCLUSIONS

In this paper for spam classification, we apply six different most popular ML classification approaches. Here we also overview the summary of these approaches. To compare the capability of different classification approaches NB, SVM, KNN, RF, LR, MLP we evaluate precision, recall. Though we also use the ROC curve and accuracy graph for performance comparison. In terms of accuracy, we find that RF achieves the highest accuracy (95%) than other approaches. However, in term of the area under curve (AUC) value SVM achieve the highest value than other approaches. Finally, we can say that for a large number of samples and feature RF and SVM perform pretty well than other approaches in the case of spam classification.

REFERENCES

[1] W.A. Awad and S.M. ELseuofi, "MACHINE LEARNING METHODS FOR SPAM E-MAIL CLASSIFICATION", International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 1, Feb 2011.

[2] Guzella, T. S. and Caminhas, W. M. "A review of machine learning approaches to Spam filtering." Expert Syst. Appl., 2009.

[3] Yuchun Tang, Sven Krasser, Yuanchen He, Weilai Yang, Dmitri Alperovitch "Support Vector Machines and Random Forests Modeling for Spam Senders Behavior Analysis" IEEE GLOBECOM, 2008.

[4] D. Puniškis, R. Laurutis and R. Dirmeikis, "An Artificial Neural Nets for Spam e-mail Recognition", *Electronics and electrical engineering*, Vol. 69, No. 5, pp. 73 – 76, 2006.

[5] Youn and Dennis McLeod, "A Comparative Study for Email Classification", *Proceedings of International Joint Conferences on Computer, Information, System Sciences and Engineering,* 2006.

[6] Cortes C, Vapnik V. Support-vector networks. Machine learning. 1995 Sep 1;20 (3):273-97.

[7] Kleinbaum DG, Dietz K, Gail M, Klein M, Klein M. Logistic regression. New York: Springer-Verlag; 2002 Aug.

[8] Donges N.," A complete guide to the random forest salgorithm", build in, 2019.