Email Spam Classification Using Machine Learning

Naman Airan and Purushottam Lal Bhari

May 5, 2022

EMAIL SPAM CLASSIFICATION USING MACHINE LEARNING

Naman Airan*, Purushottam Lal Bhari*

Department of Computer Science

Poornima Institute Of Engineering And Technology, Jaipur, Rajasthan

Email: namanairan24@gmail.com , purushottam.lal@poornima.org

Abstract- Email is a great medium of communication which is very reliable and many people use it to communicate for various purposes. Almost everyone who is part of this technical world must have at least one email account. Sometimes we receive spam mails which are tedious for us. To find these spam or not spam mails, we use Naïve Bayes, Support Vector Machine, Decision tree, and Random Forest algorithmsIt is done by finding the their precision, accuracy, recall, F-score and AUC values and comparing all the models with these known values which show which model is the best model for classification of e-mail spam.

Keywords- E-mail spam classification, Machine learning algorithms

1.INTRODUCTION

Email is the short form of electronic mail and it is defined as the exchange of information through communication channel.Emails usually come from another email address rather than entering the key board or electronic files stored on disk.Email is one of the most efficient wayfor communication with each other. The inundation of spam mails is a major problem for web users and web services these days. We also call spam mail as unwanted mail or bad mail and spam mail is the mail that the user receives without any prior information from the sender.Email spam classification save us from tedious detection takes and sometime even costly phishing scams.

2. METHODOLOGY

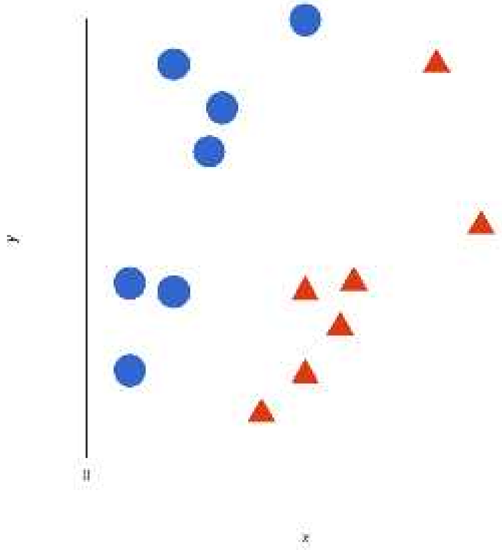
2.1Naive Bayes

Naïve Bayes algorithm is used to learns the probability of an object with certain feature belonging to a particular class and is used to determine whether the data points belong to a certain category or not. Naïve Bayes model work by correlating the use of tokens with spam and non-spam e-mails and then it uses the Base Theorem to determine the probability that the email is spam or non-spam.

2.2Support Vector Machines

SVM Modelis used to minimize classification error and maximize the margin between two classes. Support vector machines are based on the concept of decision planes that define decision boundaries.Best decision boundary is also known as hyperplane. The aim of the SVM model is to design a very fine line that separates the n-dimensional space into classes so that we can easily send new data to the right categories.

For simple visual explanation we will use two tags :red and blue with two data features: x and y and then we will train our classifier which will tell whether x/y coordinate is red or blue.



2.3 Decision Trees

Decision trees are used for classification and regression.Decision treehave tree like shapes in which each node is a decision node.A decision tree consist external node and internal node that are interlinked with each other. Decision can be made based on the internal node. If the sample of any decision tree is completely homogeneous then its entropy will be 0 and if the sample is equally divided then its entropy will be 1 and decision tree chooses only those parts whose entropy is lowest compared to the parent node and other parts. If The entropy of a model is lower then the quality of the model is good.

2.4Random Forest

A random forest algorithm collection of many decision trees. It merges Decision tree together to get more accurate and stable value. Most of the time random forest is drained with a bagging method. The begging method is dependedon the fact that the combination of learning model increases the overall result. If we combine the learning from different models and then club together then this will also increase the overall result. If the size of dataset is large then one single decision tree would lead to a same model.

3.Performance Matrices

We are using the following standard matrices to evaluate the F-score, accuracy, recall and precision by using publicly available dataset. The correctness of a classification can be evaluated by computing the number of true positives, true negatives, and false positives or false negatives.

3.1 Precision

Precision is calculated by dividing the correctly predicted positive observations to the total predicted positive observations.

Precision=TP / (TP + FP)

3.2 Accuracy

Accuracy is used to know the performance of anything and it is calculated by the ratio of correctly predicted observations to the total observations.If we have high accuracy means morethen 80% then we can say that our model is best

Accuracy = (TP+TN) / ALL

3.3 Recall

Recall is used to define how many of the true positives were found.

Recall = TP / (TP + FN)

3.4 F-Score

We have two measures Precision and Recall.F-Score is used to find the average weight of Precision and Recall. F-Score takes both false positive and false negatives into account.

3.5AUC

AUC refers to area under the ROC curve. If the closer the value of AUC is to one, the better the model is.

4. Comparison of Performance of the Models

4.1. Naive Bayes

confusion matrix and scores: [199 3]

[78 45]

Precision : 0.75Accuracy : 0.94

Recall: 0.37F-Score: 0.53

AUC: 0.67

High accuracy suggests that the model is very good at correctly classifying the mails as ham or spam. Precision value is also good at 0.75, means the model has a low false positive rate. This can be corroborated by looking at false positives found - only 3. The model has a lower recall value as compared to Decision Tree and Random Forest models. This indicates that the probability of detection is lower. Together, having a high precision and low recall means that the most of ham predictions are correct, but the model is not predicting all the ham in the test data. The AUC value is good too, as expected.

4.2. Decision Tree

confusion matrix and scores: [152 *50]*

[44 7 9]

Precision :0.71Accuracy :0.61

Recall :0.64 F-Score: 0.6

AUC: 0.64

Accuracy and Precision of the decision tree is low as compared to Naive Bayes model. However, it has a higher Recall and F-Score. If the value of recall is 0.63 that means predictions of decision tree are more complete compared to Naive Bayes and 0.71 precision define that model has low false positive rate. Since the dataset has an uneven distribution of ham and spam, F-Score becomes an important metric.It has an F-score of 0.62 which shows that this model is a very good model for precision and recall and its AUC value is also good but not as good as compared to the Naive Bayes algorithm

4.3. SVM

confusion matrix and scores: [202 0]

[123 0]

Precision : 0.62 Accuracy : 0

Recall: 0 F-Score 0

AUC 0.5

It seems like SVM is unable to predict positive class at all. if there is no predicted sample on it, it means that value of TP+FP will be 0. Hence precision and F-score are not defined and marked as 0. Accuracy is also 0 since true positive (numerator) is 0. SVM has high true negative rate and high false negative rate. The AUC score is the worst too.

4.4. Random Forest

confusion matrix and scores: [176 26]

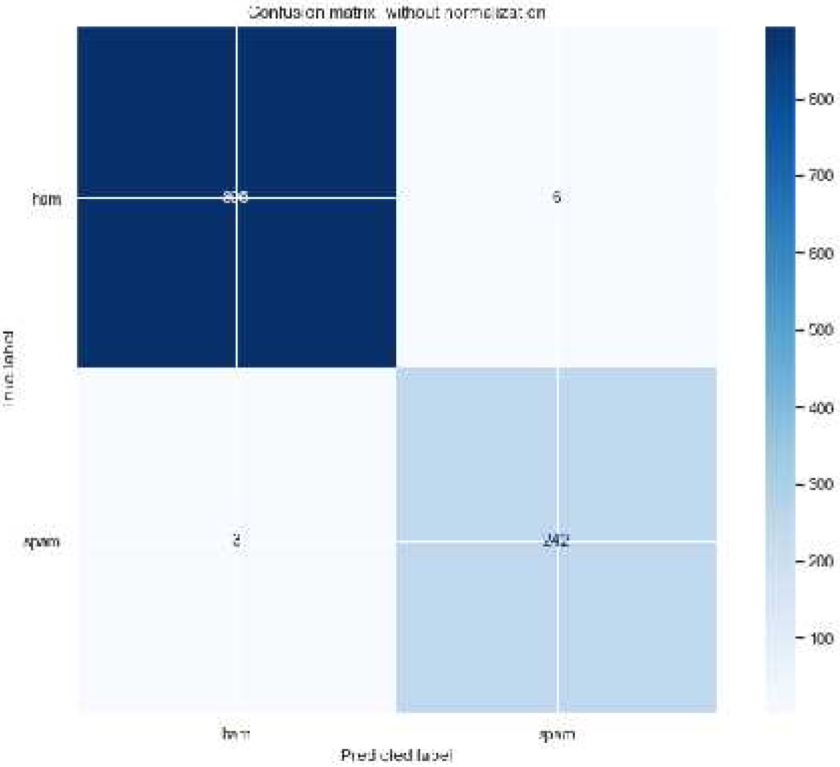
[44 79]

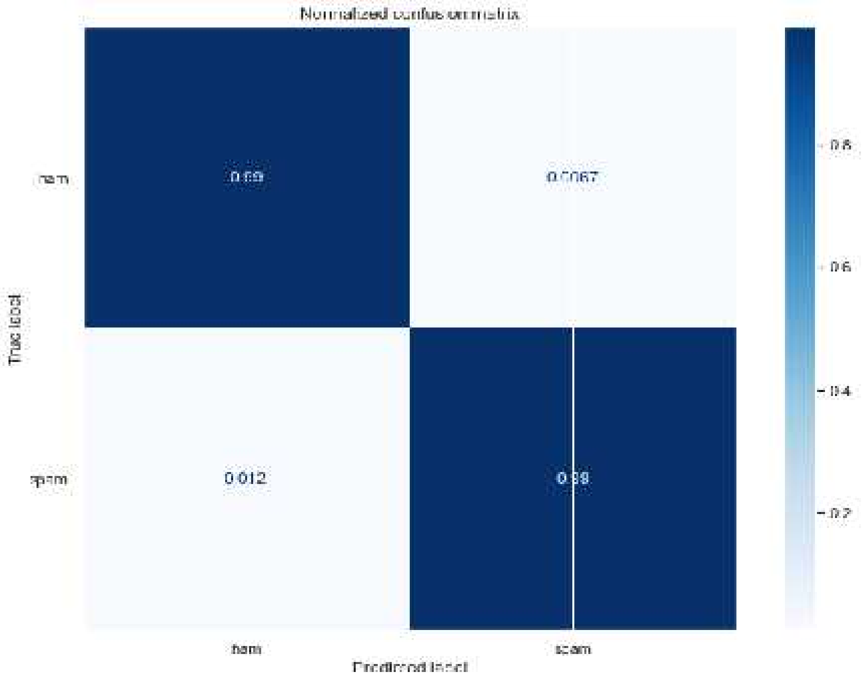
Precision : 0.78 Accuracy : 0.75

Recall : 0.64F-Score : 0.69

AUC 0.73

Random Forest algorithm had a high precision and accuracy. Considering lower accuracy as compared to Naive Bayes, good precision value indicates low false positive rate. It’s Recall value is also goodwhen compared it with Naive Bayes. Having high precision and recall suggests that the model is correctly predicting positive class (ham) and also capturing most ham in the test data. It follows that the model also has a good F-score, since it is directly proportional to Precision and Accuracy. It has best AUC score as compare too amongst all the models.





5.Results

It is clear from the comparison that SVM model did not work out very well to solve our problem of spam detection. Naive Bayes and Random Forest both algorithm are working very well. While Naive Bayes algorithm is has a high accuracy and a good precision, the recall value is poorer compared to Decision Tree and Random Forest. Since SVM model could not predict any positive values at all, its accuracy, recall and F-score were 0. As far as the F-score is concerned, Decision Tree and Random Forest have a good score as a result of good precision and recall both. Naïve Bayes and Random forest algorithm has the good AUC scores. Overall, we think that both Naive Bayes and Random Forest will be very good for spam detection.

6. CONCLUSION

In this paper we are using four types of machine learning model to classify spam or not-spam email . All the machine learning models Naive Bayes, SVM, Decision Tree and Random Forest Algorithms in detail form and comparison are available in this paper. Compared to all the machine learning models, the Naive Bayes algorithm has the highest accuracy and the SVM algorithm has the most accuracy. The Naive Bayes algorithm has an accuracy of 94% which tells whether the Naive Bayes algorithm can easily tell whether an email is spam or not. Its precision value of 75% which shows that the Naive Bayes model has a low false positive rate, due to which we can say that the Naive Bayes algorithm is the best algorithm for email spam classification.

REFERENCES

[1] M. N. Marsono, M. W. El-Kharashi, and F. Gebali, “Binary LNS-based naïve Bayes inference engine for spam control: Noise analysis and FPGA synthesis”, IET Computers & Digital Techniques, 2008

[2] Shradhanjali, Prof. Toran Verma, “E-Mail Spam Detection and Classification Using SVM and Feature Extraction”, International Journal of Advance Research, Ideas and Innovations in Technology, Rungta College of Engineering and Technology Dept. of Computer Science and Engineering Bhilai, Chhattisgarh, India.

[3]Binh T P, Indra P, Khabat K, Kamran C, Phan TT, Trinh QN, Seyed V H, Dieu T B (2018) . A comparison of Support Vector Machines and Bayesian Algorithms for Landslide Susceptibility

[4] Modeling. Geocarto International pp. 1-23. Nilam B, Namrata C, Ronit C, Shraddha M (2017). Spam E-mail detection using classifiers and Adaboost. International Journal of Computer Engineering and Application XI(VIII). https://pdfs.semanticscholar.org/c2ea/4bf0282b9b39a6ba773581332 bb0587ec4ab.pd

[5] Rizky et al. “The Effect of Best First and Spread subsample on Selection of a Feature Wrapper with Naïve Bayes Classifier for Classification of Ratio of Inpatients”. Scientific Journal of Informatics.