e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:02/Issue:09/September -2020 Impact Factor- 5.354 www.irjmets.com

EMAIL SPAM AND MALWARE DETECTION USING MACHINE LEARNING

Sudipta Ghosh*1, Subhojit Jalal*2

*1Student, department of Electronics and Communication Engineering, Amity University, Kolkata, West-Bengal, India.

*2 Student, department of Mechanical and Automation Engineering, Amity University, Kolkata, West-Bengal, India.

ABSTRACT

Spam email is one of the unwanted, unsolicited digital communication in the world of internet sent to a particular individual or a company or to a group of individuals. In the area of spam email and malware by machine learning algorithm is commonly used. The aim of this paper is to propose the machine learning algorithms: Naive Bayes, Support Vector Machines, Random Forests (Bagging) to detect the email spam. Description of the algorithms are presented and their different accuracy score is also presented in this paper. The accuracy result naïve bayes is 0.93, SVM is 0.90 and random forest is 0.97. Random forest classifier performed better than among the Decision Tree Classifier.

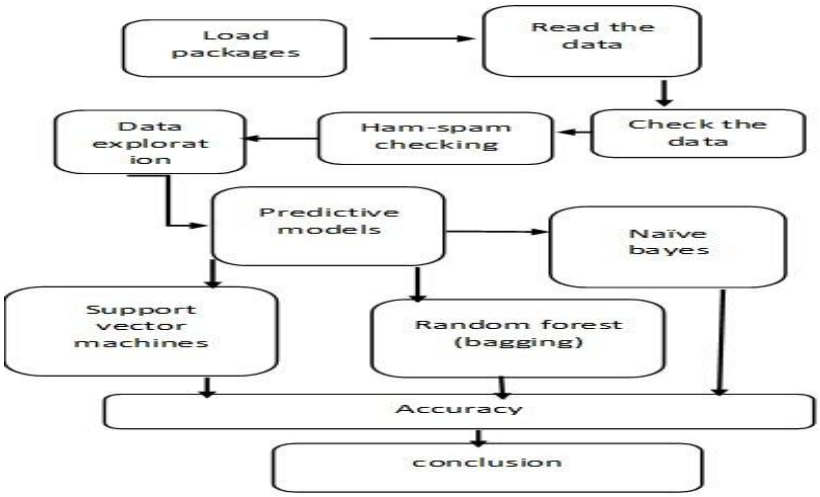
Keywords: email spam, classifier: Naïve Bayes, Support Vector Machines, Random Forest (Bagging).

I. INTRODUCTION

email system is one of the cost effective and commonly used system all over the world. Emails can be sent and received from any computer or mobile phone devices, anywhere in the world if there is any internet connection present. But day by day email system is getting threatened by spam emails which is a shotgun approach, uninvited and unwanted and unwelcomed to the receiver. spam is typically sent to a random audience or company is often characterized by misleading subject lines and poorly crafted text. It wastes the time of the receiver and it is also waste of the money of marketing department. It also damages company reputation. spam message also affects to the network capacity and usage to produce large amount of unwanted data. In recent statistics we find that around 40% of all emails included spam which about 15.4 billion email per day and that cost internet users about $355 million per year. In this paper we approach a machine learning model to detect email spam and malware in the email. Machine learning algorithms: naïve bayes, support vector machines (SVM), random-forest models are created to detect the spam emails.

We collected our dataset from Kaggle dataset, a data analysis website and started to analyze it and detect the spam emails and investigated three models. Firstly, we found the spam emails from the dataset. and then separated from the dataset then we started the prediction

II. METHODOLOGY



e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:02/Issue:09/September -2020 Impact Factor- 5.354 www.irjmets.com

1)Naïve Bayes classifier method

In this project we are classifying emails typed in by the user as either 'Spam' or 'Not Spam'. Our original dataset was a folder of 5172 text files containing the emails. We separated because this is a text-classification problem. When a spam classifier looks at an email, it searches for potential words that it has seen in the previous spam emails.

CASE 1: suppose let's take a word 'Greetings'. Say, it is present in both 'Spam' and 'Not Spam' mails.

CASE 2: Let's consider a word 'lottery'. Say, it is present in only 'Spam' mails.

CASE 3: Let's consider a word 'cheap'. Say, it is present only in spam.

If now we get a test email, and it contains all the three words mentioned above, there's high probability that it is a 'Spam' mail.

The most effective algorithm for text-classification problems is the Naive Bayes algorithm, that works on the classic Bayes' theorem. This theorem works on every individual word in the test data to make

Predictions (the conditional probability with higher probability is the predicted result).

our test email(S)is, "You have won a lottery".

P(S) = P('You') P('have') P('won') P('a') P('lottery') __ 1

Therefore, P (S |Spam) = P ('You' |Spam) P ('have' |Spam) P ('won' |Spam) P ('a '|Spam) P ('lottery' |Spam) __ 2

Same calculation for P (S |Not Spam)

If 2 > 3, then 'Spam' Else, 'Not_ Spam'.

2)Support Vector Machines

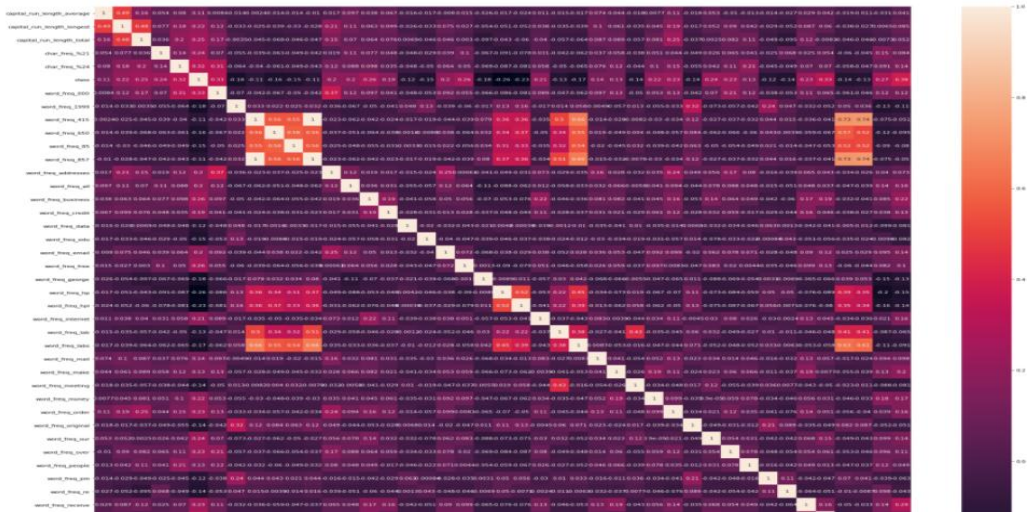
Support Vector Machine is the most sought-after algorithm for classic classification problems. SVMs work on the algorithm of Maximal Margin, i.e., to find the maximum margin or threshold between the support vectors of the two classes (in binary classification). The most effective Support vector machines are the soft maximal margin classifier, that allows one misclassification, the model starts with low bias (slightly poor performance) to ensure low variance later.

3)Random Forests (Bagging)

Random forest has nearly the same hyperparameters as a decision tree or a bagging classifier. Ensemble methods turn any feeble model into a highly powerful.

III. MODELING AND ANALYSIS

Model and Material which are used is presented in this section.



e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
Volume:02/Issue:09/September -2020 Impact Factor- 5.354 www.irjmets.com

Heatmap generation of the model.

```
mnb = MultinomialNB(alpha=1.9)          # alpha by default is 1. alpha must always be > 0.
# alpha is the '1' in the formula for Laplace Smoothing (P(words))
mnb.fit(train_x,train_y)
y_pred1 = mnb.predict(test_x)
print("Accuracy Score for Naive Bayes : ", accuracy_score(y_pred1,test_y))
```

```
Accuracy Score for Naive Bayes :  0.9381283836848216
```

naïve bayes model is working properly with 0.93 accuracy

```
svc = SVC(C=1.0,kernel='rbf',gamma='auto')
# C here is the regularization parameter. Here, L2 penalty is used(default). It is the inver
se of the strength of regularization.
# As C increases, model overfits.
# Kernel here is the radial basis function kernel.
# gamma (only used for rbf kernel) : As gamma increases, model overfits.
svc.fit(train_x,train_y)
y_pred2 = svc.predict(test_x)
print("Accuracy Score for SVC : ", accuracy_score(y_pred2,test_y))
```

```
Accuracy Score for SVC :  0.9010054137664346
```

SVM's performance is slightly poorer than

Naive Bayes

```
rfc = RandomForestClassifier(n_estimators=100,criterion='gini')
# n_estimators = No. of trees in the forest
# criterion = basis of making the decision tree split, either on gini impurity('gini'),
n infromation gain('entropy')
rfc.fit(train_x,train_y)
y_pred3 = rfc.predict(test_x)
print("Accuracy Score of Random Forest Classifier : ", accuracy_score(y_pred3,test_y))
```

```
Accuracy Score of Random Forest Classifier :  0.9760247486465584
```

IV. RESULTS AND DISCUSSION

Random Forest Classifier performs the best among the three. Decision tree classifiers are excellent classifiers. Random forest is a popular ensemble model that uses a forest of decision trees. So, obviously,

combining the accuracy of 100 trees (as estimators=100 here), will create a powerful model. displacement of all 4 cases.

The model is coming with the accuracy of 97% that we can apply to the model. In future we will try to make the model without any error that's means with 100%accuracy. this model can be used in real life scenario so that people doesn't face this problem in future.

V. CONCLUSION

Here in this paper we successfully use the machine learning algorithms ad create three models out of which the random forest classifier model is working better then two models. The model helps us to detect spam messages in the email. as the accuracy didn't come with 100%accuracy we will try to make the model with 100%accuracy as a future work.

VI. REFERENCES

[1] M. N. Marson, M. W. El-Kharosthi, and F. Gabala, "Binary LNS-based naïve Bayes inference engine for spam control: Noise analysis and FPGA synthesis", IET Computers & Digital Techniques, 2008 [2]

[2] Muhammad N. Marson, M. Wither El-Kharosthi, Fayez Gabala "Targeting spam control on middleboxes: Spam detection based on layer-3 e-mail content classification" Elsevier Computer Networks, 2009 [3]

[3] Yuchen Tang, Sven Crasser, Yuncheng He, Wailaki Yang, Dmitri Petrovitch" Support Vector Machines and Random Forests Modeling for Spam Senders Behavior Analysis" IEEE GLOBECOM, 2008

[4] Carpinteria, O. A. S., Lima, I., Assis, J. M. C., de Souza, A. C. Z., Moreira, E. M., & Pinheiro, C. A. M. "A neural model in anti-spam systems.", Lecture notes in computer science. Berlin, Springer, 2006 [9]

[5] El-Sayed M. El-Alfie, Radwan E. Abdel-Aal "Using GMDH-based networks for improved spam detection and email feature analysis "Applied Soft Computing, Volume 11, Issue 1, January 2011 [10]

[6] Li, K. and Zhong, Z., "Fast statistical spam filter by approximate classifications", In Proceedings of the Joint international Conference on Measurement and Modeling of Computer Systems. Saint Malo, France, 2006 [11]

[7] Cormack, Gordon. Smucker, Mark. Clarke, Charles " Efficient and effective spam filtering and re-ranking for large web datasets" Information Retrieval, Springer Netherlands. January 2011 [12]

[8] Analysis "Applied. Almeida, analysis "Applied, Akebi " Spam filtering: how the dimensionality reduction affects the accuracy of Naive Bayes classifiers" Journal of Internet Services and Applications, Springer London, February 2011 [13]

[9] You, S., Yang, Y., Lin, F., and Moon, I. "Mining social networks for personalized email prioritization". In Proceedings of the 15th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (Paris, France), June 28 - July 01, 2009

[10] Hao Zhang, Alexander C. Berg, Michael Maire, and Jitendra Malic. "SVM-KNN: Discriminative nearest neighbor classification for visual category recognition", IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006