A Model to Detect Spam Email Using Support Vector Classifier and Random Forest Classifier

O. E. Taylor
Rivers State University, Port Harcourt, Nigeria.
tayonate@yahoo.com

P. S. Ezekiel
Rivers State University, Port Harcourt, Nigeria.
ezekielpromise27@gmail.com

Abstract

Email spam comes in various forms, the most popular being to promote outright scams or marginally legitimate business schemes. Spam typically is used to promote access to inexpensive pharmaceutical drugs, weight loss programs, online degrees, job opportunities and online gambling. Spam is commonly used to conduct email fraud. This paper presents a model for detecting spam email using Support Vector Classifier and Random Forest Classifier. In this paper a ucl spambase dataset was trained using Support Vector Classifier and Random Forest Classifier. Random Forest Classifier had about 91.36% which is the highest accuracy while Support Vector Classifier had about 89.21% accuracy. This paper uses Random Forest Classifier in detecting spam emails, which is then saved and loaded..

Keyword: *Email, Spam, Support Vector Classifier, Random Forest Classifier*

1. Introduction

Email spam, also referred to as junk email, is an email sent without explicit consent from the recipient. Most email spam messages are commercial in nature. Whether commercial or not, many are not only annoying, but also dangerous because they may contain links that lead to phishing web sites or sites that are hosting malware or include malware as file attachments. Spammers collect email addresses from chat rooms, websites, customer lists, newsgroups, and viruses that harvest users' address books. These collected email addresses are sometimes also sold to other spammers. The use of spam has been growing in popularity since the early 1990s and is a problem faced by most email users. Recipients of spam often have had their email addresses obtained by spambots, which are automated programs that crawl the internet looking for email addresses. Spammers use spambots to create email distribution lists. A spammer typically sends an email to millions of email addresses, with the expectation that only a small number will respond or interact with the message. The term spam is derived from a famous

Monty Python sketch in which there are many repetitive iterations of the Hormel canned meat product. While the term spam was reportedly first used to refer to unwanted email as early as 1978, it gained more widespread currency in the early 1990s, as internet access became more common outside of academic and research circles.

Email spam comes in various forms, the most popular being to promote outright scams or marginally legitimate business schemes. Spam typically is used to promote access to inexpensive pharmaceutical drugs, weight loss programs, online degrees, job opportunities and online gambling. Spam is commonly used to conduct email fraud. The advance-fee scam is a well-known example -- a user receives an email with an offer that purportedly results in a reward. The fraudster presents a story where upfront monetary assistance is needed from the victim in order for the fraudster to acquire a much larger sum of money, which they would then share. Once the victim makes the payment, the fraudster will invent further fees, or stop responding. Fraudulent spam also comes in the form of phishing emails, which are emails disguised as official communication from banks, online payment processors or any other organizations a user may trust. Phishing emails typically direct recipients to a fake version of the organization's website, where the user is prompted to enter personal information, such as login and credit card details. Users should avoid opening spam emails and never respond to them or click on links in the messages. Spam email may also deliver other types of malware through file attachments or scripts, or contain links to websites hosting malware.

Spam filters can be implemented at all layers, firewalls exist in front of email server or at MTA (Mail Transfer Agent), Email Server to provide an integrated Anti-Spam and Anti-Virus solution offering complete email protection at the network perimeter level, before unwanted or potentially dangerous email reaches the network. At MDA (Mail Delivery Agent) level also spam filters can be installed as a service to all of their customers. At Email client user can have personalized spam filters that then automatically filter mail according to the chosen criteria [1]. The email has subject and body data. The following steps are required in order to apply these techniques in the filtration and classification of the emails. The first step is transferring the email contents into a numeric data. The second step is checking and identifying the similarity between the data in the header and the body of the email [2].

2. Related Works

A Survey of Email Spam Filtering Methods is explored in this section by reviewing different existing email spam filtering system regarding Machine Learning Techniques such as: Naive Bayes, SVM, K-Nearest Neighbor, Bayes Additive Regression, KNN Tree, and rules. The paper also presents the classification, evaluation and comparison of different email spam filtering system; and concluded by recommending the studying of Bayesian networks has provided a fine base for the creation of a Meta spam filter [3].

A Machine learning based spam e-mail detection presented by [4] proposed a machine learning based hybrid bagging approach by implementing two machine learning algorithms which are Naïve Bayes and Decision Tree Classifier. They divided the dataset into different sets and gave it

as input to each of the algorithm. In total, they performed three experiments and the results obtained were compared in terms of precision, recall, accuracy, f-measure, true negative rate, false positive rate and false negative rate. Two out of the three experiments were performed using Naïve Bayes and Decision Tree Classifier individually while the third experiment was the proposed system which was implemented using hybrid bagged approach. Hybrid bagged approach gave the highest accurate result of about 87.5%.

The paper "Ham and spam email classifier using machine learning techniques" classified spam emails from inboxes [5]. They applied 10 alternative classifiers on one benchmark dataset to evaluate which classifier gives a better result. A 10-fold cross validation was used to provide accuracy. Results of the classification algorithms were compared with the spam based ucl dataset. The results of the experiment shows an accuracy of about 95.45% for Random Forest Classifier compared to other classifiers used.

The paper "Email spam filtering using supervised machine learning techniques" by [1] employed supervised machine learning techniques such as Decision tree classifier, Multilayer Perceptron and Naïve Bayes Classifier to filter the email spam messages. The machine learning techniques are used in learning the features of spam emails and the model is built by training with known spam emails and legitimate emails. The results of the experiment using the supervised machine learning techniques, showed an accuracy of 98.6% for Naïve Bayes, 96.6% for Decision Tree classifier and 99.3% for Multilayer perception.

The paper "An anti-spam detection model for emails of multi-natural language" [2], investigated existing anti-spam methods. The paper highlighted some current problems and improved on an anti-spam model. They proposed a new agent-based Multi-Natural Language Anti-Spam (MNLAS) model. The Multi-Natural Language Anti-Spam model process in the spam filtering process of an email handles both visual information such as images and texts in English and Arabic languages. The Jade agent platform and Java environments are employed in the implementation of MNLAS model. The MNLAS model was tested on a 200 emails' dataset and the results showed that it was able to detect and filter various kinds of spam emails with high accuracy of about 93.32%.

The Paper "A review on different spam detection approaches" [6], discuss some approaches for spam detection. This approaches are Signatures, Mail Header Checking, Whitelist/Blacklist and Bayesian Classifier. Their discussion on this approaches are as follows: Signature is based on generating a signature having unique hash value for each spam message and the filters compare the value of previous stored values with incoming emails values. Bayesian Classifier uses posterior probability in computing all over the word in the emails. If this total value exceeds over certain threshold, then the filters will mark emails as spam. Whitelist/Blacklist approach simply creates a list. A whitelist is a list which includes the email addresses or entire domains which the user knows. An automatic white list management tool is also used by user that helps in automatically adding known addresses to the whitelist. A blacklist is the opposite of whitelist. In this list we add addresses that are harmful for users. In Mail Header Checking, they simply consist of set of rules that they match with mail headers. If a mail header matches, then it triggers the server and return mails that have empty "From" field, that have too many digits in address

that have different addresses in "To" field from same source.

The paper "Survey on e-mail spam detection using supervised approach with feature selection" [7], discuss the process of filtering the mails into spam and ham using various techniques. This technique are Machine Learning Based Technique (Support Vector Machine, Multi-Layer Perceptron, Naïve Bayes Algorithm, Decision Tree Based etc.) and Non-Machine Learning Based Technique (signature based, heuristic scanning, black and whitelist, sandboxing, mail header scanning). They concluded by saying no algorithm guarantees 100% results in spam detection but still there are some algorithms that provide high accuracy for detection of spam emails when used with feature selection technique like MLP neural network but MLP has a limitation of selecting initial information point using a randomized approach which increases the execution and model building time of the MLP algorithm.

The paper "An efficient spam filtering techniques for email account" [8], presented an efficient spam filter technique to spam email based on Naive Bayes Classifier. They collected a statistical data which they used in training the Bayesian Classifier. This Bayesian filtering works by evaluating the probability of different words appearing in legitimate and spam mails and then classifying them based on those probabilities.

The paper "An efficient spam filtering using supervised machine learning techniques" [9], employed a supervised machine learning techniques to filter the email spam messages. The supervised machine learning techniques used are C 4.5 Decision tree classifier, Multilayer Perceptron and Naïve Bayes. They used Naïve Bayes Classifier for learning the features of spam emails and the model is built by training the mentioned Classifiers with known spam emails and legitimate emails. They came up with a predicted accuracy, Naïve Bayes Classifier 98.6%, C 4.5 Decision tree classifier 96.6% and Multilayer Perceptron 99.3%

The paper "A survey of existing e-mail spam filtering methods considering machine learning techniques" [10], illustrates a survey of different existing email spam filtering system regarding Machine Learning Technique (MLT) such as Naive Bayes, SVM, K-Nearest Neighbor, Bayes Additive Regression, KNN Tree, and rules. However, here they present the classification, evaluation and comparison of different email spam filtering system and summarize the overall scenario regarding accuracy rate of different existing approaches.

3. Design Methodology

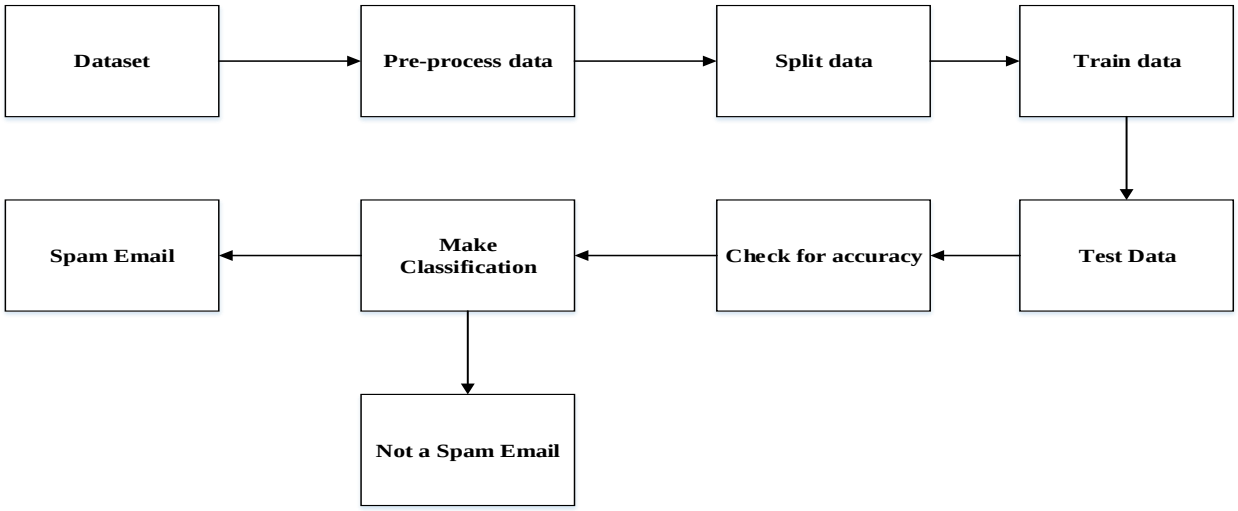

Figure 1: Architecture of the proposed system

This system uses a University Collection London (UCL) spambase dataset which was created by Mark Hopkins, Erik Reeber, George Forman and Jaap Suermondt. The dataset contains 58 columns in which the last column denotes whether an email was spam or not. This dataset was being preprocessed using min_max_scaler, making sure that all the values are properly scaled for efficient result. After cleaning, the dataset was split into x and y variables where x variable contains 57 columns whereas y variable contains class column which indicate if an email is spam or not a spam. After splitting the dataset, the dataset was being trained using two machine algorithms which are Support Vector Classifier and Random Forest Classifier. The dataset was tested based on accuracy on this two machine algorithm. After checking for accuracy, the model was saved and used for making classification to tell when an email is a spam one and also not a spam one.

	word_freq_make	word_freq_address	word_freq_all	word_freq_3d	word_freq_our	word_freq_over	word_freq_remove	word_freq_internet	word_freq_order
0	0.00	0.64	0.64	0.0	0.32	0.00	0.00	0.00	0.00
1	0.21	0.28	0.50	0.0	0.14	0.28	0.21	0.07	0.00
2	0.06	0.00	0.71	0.0	1.23	0.19	0.19	0.12	0.64
3	0.00	0.00	0.00	0.0	0.63	0.00	0.31	0.63	0.31
4	0.00	0.00	0.00	0.0	0.63	0.00	0.31	0.63	0.31

5 rows × 58 columns

word_freq_mail	...	char_freq_%3B	char_freq_%28	char_freq_%5B	char_freq_%21	char_freq_%24	char_freq_%23	capital_run_length_average
0.00	...	0.00	0.000	0.0	0.778	0.000	0.000	3.756
0.94	...	0.00	0.132	0.0	0.372	0.180	0.048	5.114
0.25	...	0.01	0.143	0.0	0.276	0.184	0.010	9.821
0.63	...	0.00	0.137	0.0	0.137	0.000	0.000	3.537
0.63	...	0.00	0.135	0.0	0.135	0.000	0.000	3.537

Figure 2: showing some information about the dataset for the first 5 rows

4. Result and Discussion

In this paper, a machine learning model was being trained to detect if an email is a spam email or not. This model uses a spambase dataset which have 58 columns. The dataset was being cleaned and processed making sure that there are no null values present. The values of the dataset were well scaled using min_max_scaler for proper fitting in training of the model using the two machine learning algorithms. The dataset was further divided into x and y variables. Where x variable contains 58 columns (informations of spam and real emails) and the y variable contains the output. The x and y variable were further divided into x_train, x_test, y_train, y_test. This x_train, y_train were being fitted or trained using two machine learning algorithms which are Support Vector Classifier and Random Forest Classifier. These two machine learning algorithms were tested for accuracy. The Support Vector Classifier came up with an accurate result of about 89.21% when kernel = 1 while Random Forest Classifier came up with an accurate result of about 95.36% where number of estimator = 2. After checking for accuracy, Random forest Classifier had the highest number of accurate result which is 95.36%. Random Forest Classifier was then saved and used in checking for spam emails.

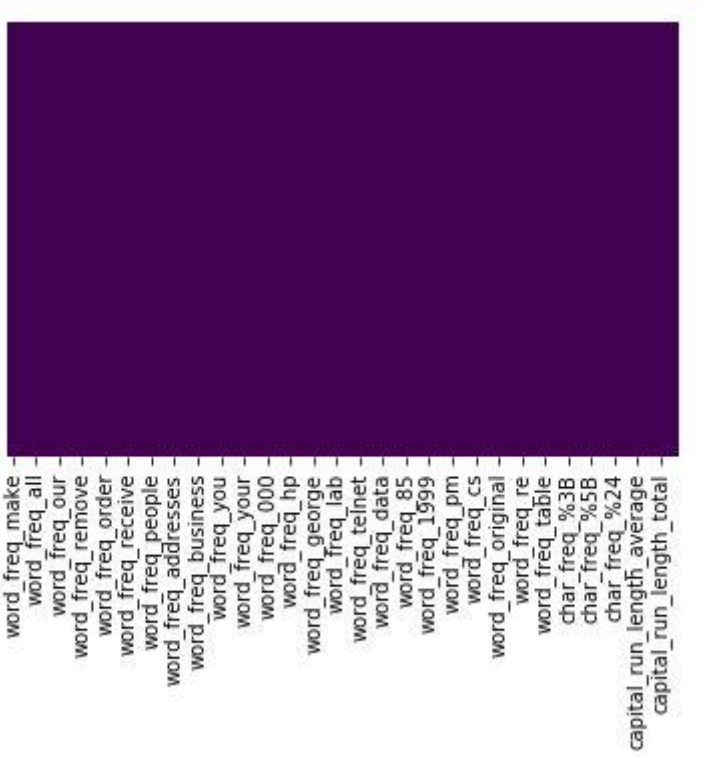

Figure 3: showing that the dataset has been clean (no null values are present)

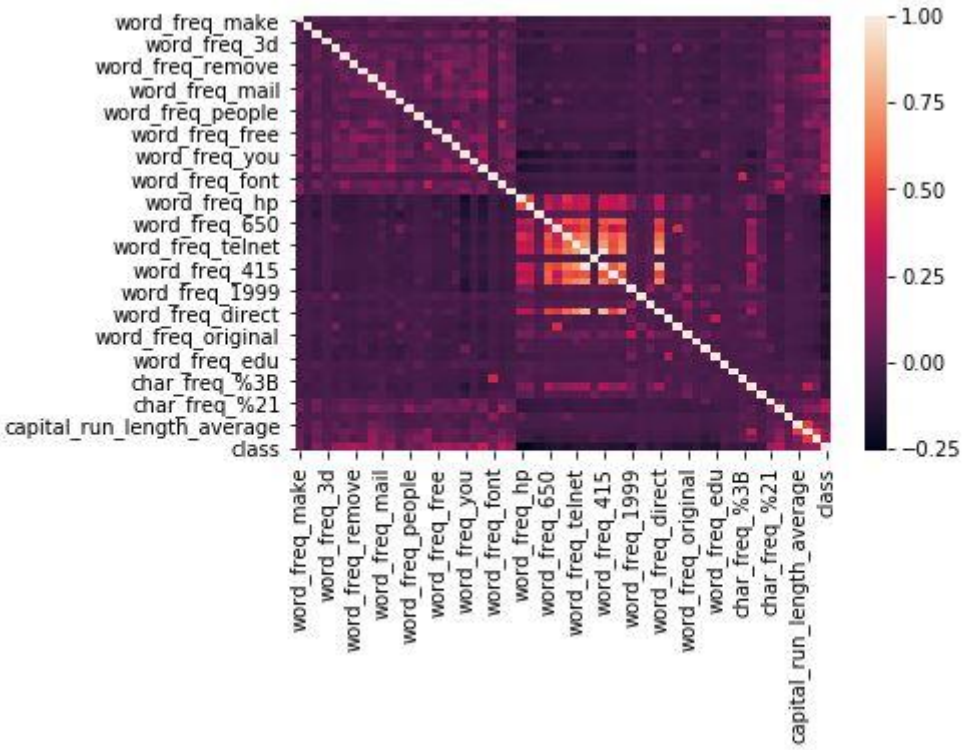

Figure 4: showing a correction matrix of the dataset

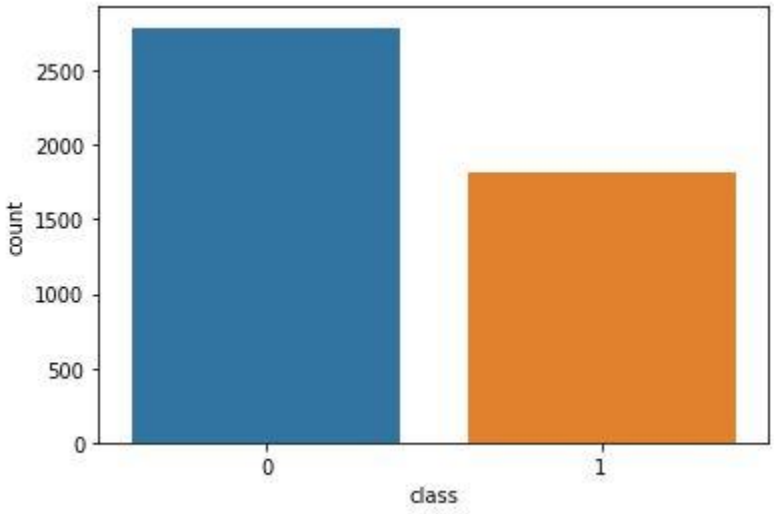

Figure 5: showing a count plot of real and spam emails, were 1 represents spam emails and 0 represents real email

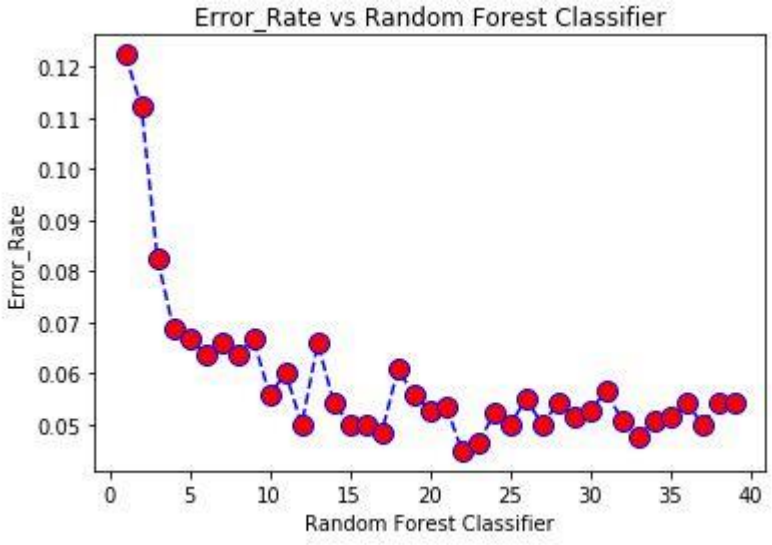

Figure 6: showing error rate vs random forest classifier

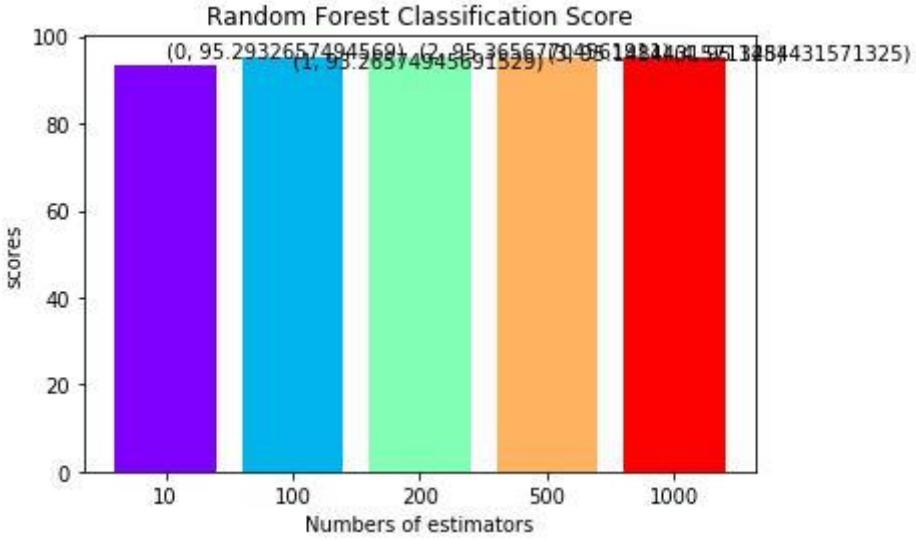

Figure 7: showing accurate results as n goes from 0 to 4

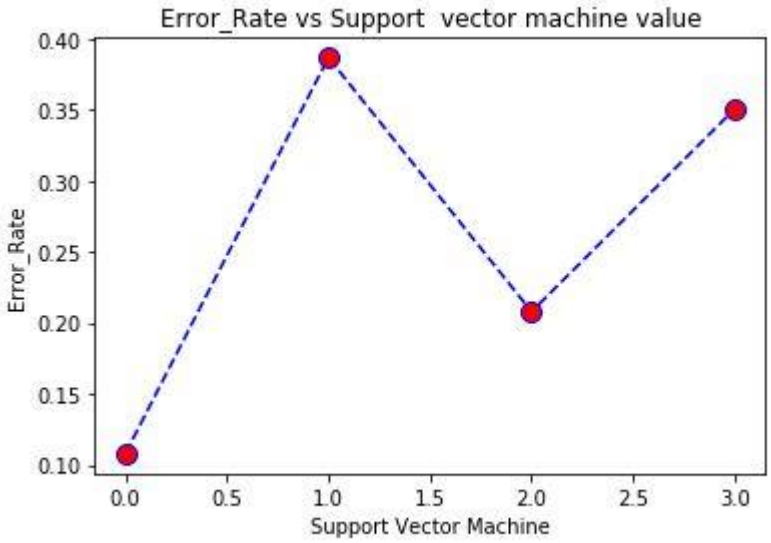

Figure 8: showing error rate vs support vector machine

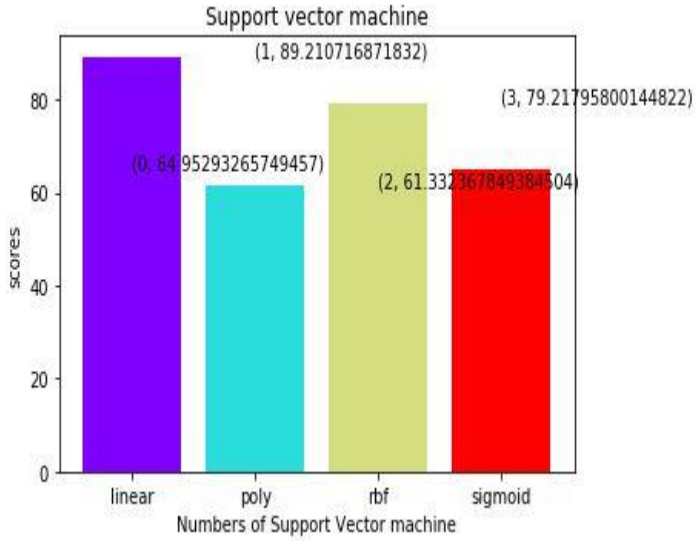

Figure 9: showing accurate results for support vector machine as n goes from 0 to 3

5. Conclusion and future scope

This paper presents two machine learning algorithms (Support Vector classifier and Random Forest Classifier) which are used in training and analyzing a machine learning model for

detecting spam emails. A dataset which contains 58 columns was used in training the machine model. After training and testing for accuracy, Support Vector Classifier came up with an accurate result of about 89.21% when kernel equals 1 while Random Forest Classifier came up with an accurate result of about 91.36% where number of estimator = 2. This paper can further be extended by comparing the accuracy and performance of other machine learning classifiers like, Naïve Bayes K-Nearest Neighbor, Logistic Regression, Linear Regression. It can further be extended using Keras and Tensorflow in training the network.

References

[1]. V.Christina, S. Karpagavalli, G. Suganya, "Email Spam Filtering using Supervised Machine Learning Techniques" International Journal on Computer Science and Engineering 2(9), pp. 3126-3129, 2010.

[2]. M. A. Mohammed, S. A. Mostafa, O. I. Obaid, S. R. Zeebaree, M. K. Ghani, A. Mustapha, M. F. Fudzee, M. A. Jubair, M. H. Hassan, A. Ismail, D. A. Ibrahim. "An anti-spam detection model for emails of multi-natural language". Journal of Southwest Jiaotong University. 54(3), pp. 0258-2724, 2019.

[3]. M. Sharma, S. Sharma, "A Survey of Email Spam Filtering Methods" Control Theory and Informatics 7, 2224-5774, 2018.

[4]. P. Sharma, U. Bhardwaj, "Machine Learning Based Spam Email Detection" International Journal of Intelligent Engineering and System 11(3), pp. 1-10, 2018

[5]. M. Bassiouni, M. Ali, E. A. El-Dahshan. "Ham and Spam Email Classification Using Machine Learning Techniques" Journal of Applied Security Research 13(8), pp. 315-331 2018.

[6]. S. N. Rekha, "A Review on Different Spam Detection Approaches" International Journal of Engineering Trends and Technology (IJETT). 11 (6), pp. 315-318, 2014.

[7]. H. Kaur, P. Verma. "Survey On E-Mail Spam Detection Using Supervised Approach With Feature Selection" International Journal of Engineering Sciences & Research Technology 6(4), pp.2277-9655. 2017.

[8]. S. Roy, A. Patra, S.Sau, K.Mandal, S. Kunar. "An Efficient Spam Filtering Techniques for Email Account" American Journal of Engineering Research (AJER). 2(10) pp. 63-73, 2013.

[9]. D. Mallampati. "An Efficient Spam Filtering using Supervised Machine Learning Techniques" International Journal of Scientific Research in Computer Science and Engineering. 6(2), pp.33-37, 2018.

[10]. H. Bhuiyan, A. Ashiquzzaman, T. I. Juthi, S. Biswas. "A Survey of Existing E-Mail Spam Filtering Methods Considering Machine Learning Techniques". Global Journal of Computer Science and Technology: C Software & Data Engineering. 11(2). pp. 0975-4172, 2018.