Engineering Universe for scientific Research and management
ISSN (Online): 2319-3069
Vol. 11 - Issue VIII
August 2019

A MACHINE LEARNING MODEL OF EMAIL CLASSIFICATION FOR SPAM DETECTION

Garima Yadav[1], Prof. Chinmay Bhatt[2], Dr. Varsha Namdeo[3]
[1,2,3] Department of CSE , RKDFIST, Bhopal, India
garima.yadav.1007@gmail.com

ABSTRACT:

Nowadays email is an individual most prominent quickest and least expensive method for communications. It has become a bit of everyday life for many residents for their data sharing. Because of its simplicity email is presented to a great deal of dangers. A standout amongst the most basic dangers toward email is spam: at all spontaneous beneficial correspondence. The development of spam movement is proper a stressing issue given that it expends the system data transfer capacity, time of clients and squanders memory and causes monetary misfortune together the clients with the associations. In this paper, we recommend a model that performs characteristic choice for remorseless spam recognition in email among the goal of streamlining the grouping parameter, the figure precision and computation time for later on characterization calculations. A work of ling spam dataset will be utilize for the technique of feature determination, and after that the arrangement of picked features will be validate with four classifiers: Support Vector Machine (SVM), Naïve Bayes (NB), Logistic Regression and Random Forest.

Keywords: SVM, E-mail, Naive Bayes, Classification, Spam, Malicious.

1. INTRODUCTION

Currently, Electronic mails contain become one of the most significant way of communication. Unfortunately, as the significance of messaging enlarges, the quantity of spam messages send to users moreover increases. In detail, spam emails are the same messages send to a lot of users. Spam emails have unusual functions. A few of them provide for promotion issues, others are conscientious of dispersal computer viruses as well as there be present spam messages intended to appropriate the user economic identities [1]. As of late unplanned business/mass email or else called spam, turn into a most important difficulty over the web. Spam is exercise in futility, storage space and correspondence data transmission. The issue of spam email has been expanding for a considerable extent of time. In late insights, 40% of all messages are spam which around 15.4 billion email for every day and that cost web clients about $355 million every year. Programmed email filtering is by all accounts the best technique for countering spam right now and a tight rivalry amongst spammers and spam-filtering strategies is removal on. Just quite a long, while back the vast majority of the spam could be dependably managed by blocking messages originating from specific locations or filtering out messages with certain subject lines. Spammers started to utilize a few precarious techniques to conquer the filtering strategies like utilizing irregular sender addresses as well as attach arbitrary characters to the start or the finish of the message subject line [2, 4].

Information assembles along with machine learning are the two universal methodologies utilized as an element of email filtering. In learning designing methodology an arrangement of guidelines must be determined by which messages are ordered as spam or ham. An arrangement of such standards ought to be made either by the client of the filter, or by some other specialist (e.g. the product organization that give demonstrates on the grounds that the standards must be continually refreshed and kept up, which is an exercise in futility and it isn't advantageous for generally clients. Machine learning approach is more effective than information building approach; it doesn't require indicating any principles [4, 5]. Rather, an arrangement of preparing tests, these examples is an arrangement of pre ordered email messages. A particular calculation is then used to take in the classification rules from these email messages. Machine learning approach has been generally examined and there are lots of algorithms can be utilized as a part of email filtering. In this paper, an individual email record of more than standard amount of messages is used for examination and valuation. The focal point is to learn the email substance and

address along with classify all email into one of two: private, professional with other based on content, sender, as well as heading. Methodologies a deal of strategy have to be determined through which messages are prepared as spam or else ham.

This rest of this paper is prepared as follows. Section 2 reviews the broad literature that uses machine learning in E-Mail Classification, as well as a few of the explanations of two methods. Section 3 describes the literature review of various machine learning methods f used in classifications. Section 4 describes proposed work in which we explain algorithms for e-mail classification and its flow chart, and Section 5 presents the results of the analysis and discusses the relative algorithms for e-mail classification along with we represent them using confusion matrix as well as its accuracy. Section 6 concludes the paper by discussing the implications of the results for e- mail classification and analysis, and also explains future scope of this work.

2. MACHINE LEARNING IN E-MAIL CLASSIFICATION

Machine taking in field is a subfield from the expansive field of artificial intelligence, these plans to make machines ready to learn like human. Learning here means comprehended, observe and converse to data about some statistical phenomenon. In unsupervised learning one tries to reveal shrouded regularities (bunches) or to identify anomalies in the information like spam messages or system interruption. In email filtering errand a few features could be the pack of words or the subject line analysis.[7] Thus, the contribution to email classification assignment can be seen as a two dimensional matrix, whose axes are the messages along with the features. Email classification assignments are frequently separated into a few sub-undertakings. To start with, Data accumulation and representation are for the most part issue particular (i.e. email messages), second, email feature choice and feature diminishment endeavor to decrease the dimensionality (i.e. the quantity of features) for the rest of the means of the task. At long last, the email classification period of the procedure finds the genuine mapping between training.

2.1 Naive Bayes Classifier:

During 1998 the Naïve Bayes classifier (figure 1) was planned for spam identification. Bayesian classifier is functioning on the dependent events along with the probability of an occurrence happening inside the future that can be detect from the earlier occurring of the similar event [9]. This technique can be use to categorize spam e-mails; words probabilities play the major rule at this time. If a few words happen often in spam but not within ham, then this incoming e-mail is possibly spammed. Naive bayes classifier technique has happen to a really popular method during mail filtering software. Bayesian filter ought to be trained toward work efficiently. Every word has definite likelihood of occurring within spam or ham email inside its database. If the whole of words probability exceeds a confident limit, the filter will mark the e-mail to moreover category.

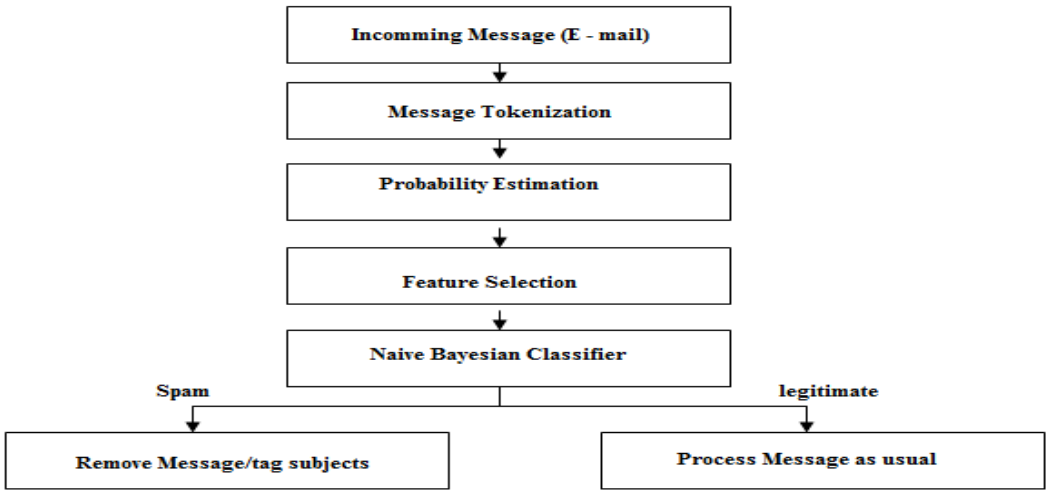

Figure 1: Naive Bayes Classifier

Now, only two categories are necessary: spam or else ham. Approximately every one the statistic-based spam filters exercise Bayesian probability calculation toward join individual token's statistics to a general

score [1], furthermore make filtering decision base lying on the score. The statistic we are commonly concerned for a symbol T is its spamminess (spam rating) [8], consider as follow:

$$S[T] = \frac{C_{Spam}(T)}{C_{Spam}(T) + C_{Ham}(T)}$$

Where CSpam (T) as well as Cham (T) are the amount of spam or else ham messages contain token T, correspondingly. To compute the likelihood for a message M among tokens $\{T_1,.....,T_N\}$, one requests to merge the individual token's spamminess toward estimate the generally message spamminess.

2.2. Support Vector Machine

Electronic mail is an input revolution attractive place more than conservative communication systems appropriate to its convenient, fast, easy, along with economical, to utilize nature. A major block surrounded by electronic communications is the huge allocation of unwanted, risky emails known like spam emails. A key apprehension is the rising of appropriate filters that can adequately confine those emails as well as get elevated performance rate. Machine learning (ML) researchers contain developed many approaches within order to deal by this difficulty. Within the structure of machine learning, support vector machines (SVM) have ready a great part to the growth of spam email filtering. Base on Support Vector Machine, dissimilar method have been intended during text classification approaches (TC). A serious problem while by SVM is the choice of kernels as they explicitly affect the panel of emails in the value space [10]. Now figure 2; explain the spam filtering by SVM.

2.3. Logistic Regression

The logistic regression model associations the probability of an email life form spam (πi) toward the prediction variables (x1i;… … ..; xij) through a framework very like that of multiple regression. Since the response is binary, we need to locate a suitable change in order to make the regression model work. A characteristic change for πi is the logistic change:

$$\text{logit}(\pi_i) = \ln\frac{\pi i}{(1-\pi i)} \quad (1)$$

The logistic regression model is specified by:

$$\ln\frac{\pi i}{(1-\pi i)} = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots\ldots + \beta_j x_{ij} \quad (2)$$

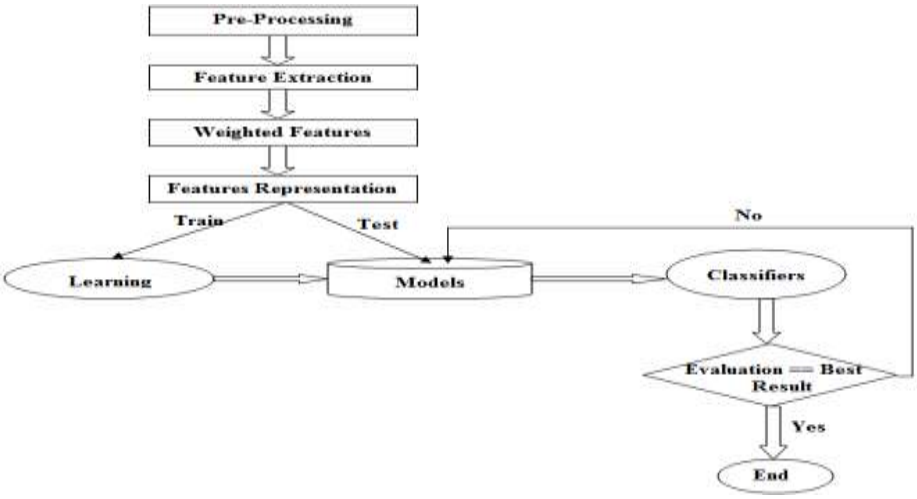

Figure 2: SVM Classifier

Note that because the probability of an email individual spam (π_i) is a numeral between zero as well as one, the $\log(\frac{\pi i}{(1-\pi i)})$ can take on a few real number:

$$0 \le \pi_i \le 1 = -\infty < \ln\frac{\pi i}{(1-\pi i)} < +\infty$$

The relation among $P(Y_i = 1)$ is obtain through solving 2 for π_i. We obtain:

$$P(Y = 1 \| X = x) = \pi_i = \frac{\exp(\alpha + \beta 1xi1 + \beta 2xi2 + \cdots\ldots + \beta jxij)}{\exp(1 + \alpha + \beta 1xi1 + \beta 2xi2 + \cdots\ldots + \beta jxij)} \quad (3)$$

$$= \frac{\exp(\alpha + \beta xj)}{1 + \exp(\alpha + \beta x)}$$

Equation 3 is the logisic regression form that will be utilize all through this paper.

We describe the odds as

$$\Omega = \frac{\pi i}{1-\pi i} \quad (4)$$

where the odds is the likelihood of the outcome spam divided with the likelihood of the outcome no spam.

By taking the logarithm on the two sides we get equation 2. The logistic regression coefficient corresponds to the change in the long odds, for each variable respectively. The exponentiated types of the coefficients correspond to the odds proportion [21].

3. LITERATURE SURVEY

There are some research works that apply machine learning techniques in email classification; showed that the navie Bayes email content classification could be adjusted for layer-3 preparing, without the requirement for reassembly. Recommendations on pre-identifying email parcels on spam control center boxes to support opportune spam detection at getting email servers were introduced.

W. El-Kharashi, et al.[9] They introduced equipment engineering of naive Bayes inference motor for spam control utilizing two class email classification. That can order more 117 millions features for every second given a stream of probabilities as information sources. This work can be reached out to examine proactive spam taking care of plans on accepting email servers and spam throttling on network gateways.

Y. Tang, S. Krasser, et al. [10] a framework that utilized the SVM for classification reason, such framework remove email sender conduct information in light of worldwide sending dispersion, investigate them and allot an estimation of trust to every IP address sending email message, the Experimental outcomes demonstrate that the SVM classifier is viable, precise and substantially speedier than the Random Forests (RF) Classifier.

Yoo, S., Yang, Y., et al. [11] created personalized email prioritization (PEP) strategy that uniquely center around investigation of individual informal communities to catch client gatherings and to acquire rich features that speak to the social parts from the perspective of specific client, and also they built up a regulated classification structure for modeling individual needs finished email messages, and for anticipating significance levels for new messages.

Guzella, Mota-Santos , et al. [13], an immune-enlivened representation, named innate along with adaptive artificial immune system (IA-AIS) as well as associated to the issue of identifiable evidence of spontaneous mass email messages (SPAM). It coordinates substances similar to macrophages, B and T lymphocytes, modeling both the innate and the adaptive immune systems. An execution of the calculation was fit for distinguishing over 99% of genuine or SPAM messages specifically parameter designs. It was contrasted with an improved variant of the naive Bayes classifier, which have been accomplished to a great degree high right classification rates. It has been reasoned that IA-AIS has a more noteworthy capacity to distinguish SPAM messages, despite the fact that the recognizable proof of honest to goodness messages isn't as high as that of the executed naive Bayes classifier.

Webb et al.' [14] web spam with how to apply email spam detection techniques to identify spam web pages. Alike to the way to deal with identify spam in emails, web pages are examine for specific features that may categorize them as spam pages, for example, utilizing keywords stuffing, unrelated famous words, etc. [12] represents one more instance of web or else connect spam research paper. Open networks, Blogs, news or else even e-commerce websites these days permit users to concern their comments or else feedback. Spammers utilize such capability to post spam messages among those posts. Therefore spam detection techniques must be additionally used to permit programmed detection of such posts.

Sculley and Wachman [15] examine too calculations, for example, VSM for email, web, and blogs and web and connection spam recognition. The substance of the email or else the web page is analyzed by disparate regular language processing methodology, for example, NGram, Bags of words, etc. The effect of an exchange parameter in VSM is evaluated utilizing divergent setting value intended for such parameter.

Zhou et al. [16] spam-based categorization scheme of three category. In adding to exemplary spam and not spam category, a third uncertain category is provide to extra flexibility to the prediction calculation. Undecided emails ought to be re-examined and collect more data to be capable then to pundit whether they are spam or else not. Xie et al. [17] attempt to whole up features that can recognize spam intermediary that are used to toss a huge number of spam emails. Creators take a gander at network interrelated behaviors that can most likely identify such spam intermediary. [16][17] Evaluate apply uneven set on spam recognition with unique rule execution scheme to get the best coordinating one. UCI Spam base is use in the investigational examine (machine learning repository or repository). Ozcaglar[18]. Unlike papers discussed the utilizing of special calculations and likewise apply the calculations in special places between email senders alongside receivers.

Carmona-Cejudo et al.'s [19], real time email category as well as introduce GNUs mail open source use designed for email file classification. The application be developed to parse emails from unique email clients alongside perform several information mining investigation with WIKA information mining instrument. In email database categorization is likewise base lying on the time of email messages.

4. PROPOSED WORK

In this Research paper we have proposed a machine learning method for categorization computations that obtain by attempt at data analysis. We took ling spam corpus dataset for the purpose of experiment, which is incredibly enormous dataset and it includes various emails and these emails are categorized to organize emails and analyze the emails, which is enlighten through figure 3. Here now we initially describe regarding how ling spam corpus dataset preprocessed step by step.

4.1 Ling Spam corpus dataset

Initially categorization computation is compared based on confusion matrix and precision. These computations of categorization have been employed on the ling spam corpus dataset, which contains of enormous emails for the purpose of training and testing. The step by step process involved in this are as following:

- Initially organize the dataset by preprocessing of it
- Lexicon table will be formed for all word
- Do Feature extraction
- Train the classifier and then testing

1. Organize the dataset

During this procedure we have to preprocess the dataset. For this purpose we have acquired the ling corpus dataset which consist of 702 training emails and 260 testing emails, so we have collectively approximately 962 emails.

a) **Stop Words Removal –** Stop words like “the”, “and", "of", and so on are very common words in English sentences over and above they aren’t really significant for detecting ham or spam emails thus these words have been separate initially from the emails.

b) **Stemming –** during this process, compile mutually the different altered types of a word so they could be scrutinize as a meticulous article. For instance, "include", "includes," and "included" would all be symbolized as "include". The circumstance of the sentence is similarly conserved in lemmatization as contrasting to stemming (another term in text mining which doesn’t consider denotation of the sentence).

2. Lexicon table Formation

Generally the main line of the email is subject and the third line includes the main body of the email. We now carry out word analysis on the content to identify the spam emails. As primary step, we must produce a lexicon of words and their occurrence. For this work, training of 700 emails is utilized.

Once the lexicon is produced we could embrace only a small number of lines of code to the above ability to eliminate non words.

3. Feature extraction

Once the lexicon is organized; we could acquire word count vector of 3000 dimensions proposed for all email of training set. All word ensured that the vector hold the frequency of 3000 words in the training. Obviously we might have estimated at this point a vast segment of them determine be zero. For example, suppose we have 500 words in our lexicon. All words verify the vector enfold the frequency of 500 lexicon words in the training file. Suppose text within training was "Get the work done, work done" then it will be pre assembled as:

[0,0,0,0,0,… … .0,0,2,0,0,0,… … ,0,0,1,0,0,… 0,0,1,0,0,… … 2,0,0,0,0,0]

Here, each one of the word counts are placed at 296^{th}, 359^{th}, 415^{th}, 495^{th} catalog of 500 length word count vector additionally to the remaining will zero.

4. Classifiers Training

We trained the dataset by four machine learning algorithms i.e. Naive Bayes, Support Vector Machines, Logistic Regression and Random Forest.

- The Logistic Regression is renowned machines learning algorithm for twofold classification. It acquires actual weighted sources of data and constructs a prediction with consideration to the probability of the data residing to its default class.
- Naive Bayes is a conventional classifier with enormously renowned technique proposed for text mining problem. It is a supervised probabilistic classifier based on Bayes theorem cooperative between every match of characteristics.
- The decision trees night different by limiting the attributes that the greedy computation could appraise at each split instant that creating the tree. This is known as the Random Forest computation.
- SVMs are supervised paired classifiers which are incredibly efficient when you have superior number of attributes. The purpose of SVM is to split a few subset of training data from remaining dataset is known as the support vectors. The decision capability of SVM model that forecasts the group of the test data is based on support vectors and formulates utilization of a kernel.

Once the classifiers are trained, we could test the performance of the proposed models on test dataset.

We eliminate word comprise vector matrix for all email test dataset with forecasting its group i.e. ham or else spam, through the various machine learning algorithms NB classifier, Logistic regression, SVM model, Random forest.

4.2. Proposed Algorithm

Test-set include 130 spam emails as well as 130 non-spam emails. If you have approach so far, you will discover below result. I have exposed the confusion matrix of the test-set intended for together the models. The diagonal parts represent the accurately known (true identification) mails where as non-diagonal element represents incorrect classification (false identification) of mails.

Algorithm:
Input: Ling Corpus Data set
Output: Result every of the Classifier
1. Download the dataset commencing website or else from inherent library.
2. Preprocess the data set by preprocessing method
3. Subsequent to preprocessing relate a variety of machine learning classification algorithm similar to naïve bayes, logistic regression, random forest and SVM lying on the preprocess dataset.
4. Work out the accuracy of the classification technique.
5. Evaluate every one of the classification technique.

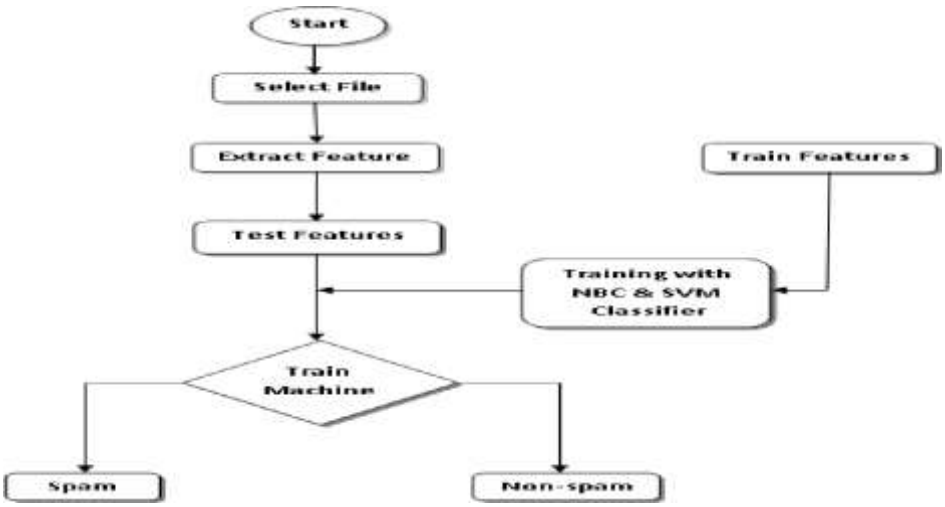

Figure 3: proposed method flow graph for classification via NBC & SVM

5. RESULT ANALYSIS AND ITS PARAMETERS METRICS USED

Here, we use the Python version 3.6 for examination as well as its parameter which is use of this investigation. The series of steps and every one of the computations with it will be showed in this segment, in both parallel and sequential evaluation. The best four models for evaluations are likewise presented here.

5.1. Confusion Matrix:

A confusion matrix is an abstract of prediction outcome on a classification difficulty.
The numeral of correct along with wrong predictions are sum up with count values as well as broken down through every class. This is the key in to the confusion matrix. The confusion matrix demonstrates the method in which your classification model is confused while it makes predictions. It gives us approaching not only into the errors individual made through a classifier but further prominently the kind of errors that are being made.

	Set 1 Predicted	**Set 2 Predicted**
Set 1 Actual	TP	FN
Set 2 Actual	FP	TN

Here,

Set 1: Positive

Set 2: Negative

Explanation of the Terms:

Positive (P): Examination is positive (for case: is an apple).

Negative (N): Examination is not positive (for case: is not an apple).

True Positive (TP): Examination is positive, along with is predicted to be positive.

False Negative (FN): Examination is positive, other than is predicted negative.

True Negative (TN): Examination is negative, along with is predicted to be negative.

False Positive (FP): Examination is negative, other than is predicted positive.

5.2. Classification Accuracy:

Classification Accuracy is known through the relation: Though, there are problems through accuracy. It assumes equivalent costs for mutually type of errors. A99% accuracy may be excellent, good, middling, poor or else awful depending leading the problem.

$$\textbf{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Here, we have use classification algorithms available within particular library. Initially, we will estimate the confusion matrix subsequent that we will calculate the accuracy during by function or confusion metrics. accuracy_score; firstly we will demonstrate the output of Ling Spam Dataset which is specified here:

According to Naive Bayes

	Ham	**Spam**
Ham	129	1
Spam	9	121

Accuracy Score: 96.1538461538%

According to SVM demonstrate the confusion matrix

	Ham	**Spam**
Ham	126	4
Spam	6	124

Accuracy Score: 96.1538461538%

According to logistic regression, we will illustrate the output of Ling Spam Dataset which is given below:

	Ham	**Spam**
Ham	126	4
Spam	1	129

Accuracy Score: 98.0769230769%

As indicated by Random forest

	Ham	**Spam**
Ham	124	6
Spam	6	124

Accuracy Score: 95.3846153846%

Support Vector Machine, Naïve Bayes, Logistic Regression and Random Forest classifier were implemented and compared to each other in terms of accuracy score. The comparison of classifiers results are shown in the following table.

Table 1: Comparisons of previous and present result on given data set

Method	**Base Methods**	**Data Set (Ling-Spam Corpus)**
SVM	91 %	96.15 %
Naive Bayes	92 %	96.15 %
Logistic Regression	-	98.07 %

Comparative study of based methods to be used in previous paper and proposed methods in given table 1:

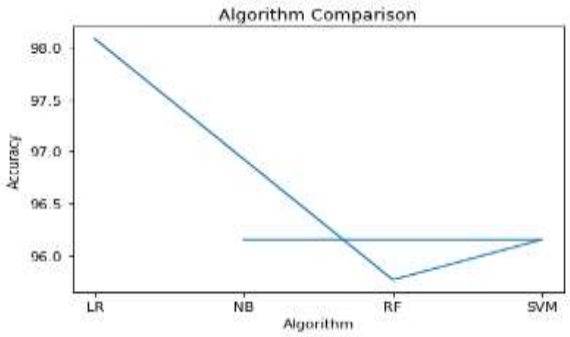

Figure 4: A Comparative Study of different classifier on ling_spam_corpus Dataset

Accurate classification results for every classification methods lying on ling_spam_corpus Dataset as well as comparisons to all other with accuracy are shown within Figure 4.

6. CONCLUSION

In this paper we review some of the most prominent machine learning methods and of their relevance to the problem of spam e-mail classification. Descriptions of the calculations are presented, and the correlation of their performance on the Ling corpus Spam Dataset is presented, the experiment demonstrating a very encouraging results specially in the calculations that isn't well known in the commercial e-mail filtering packages, spam recall percentage in the five methods has the accuracy values, while in term of accuracy we can find that the Naïve bayes and SVM methods and Logistic Regression methods has a very fulfilling performance amongst the other technique, more research should be done to rise the performance of the Naïve bayes either through hybrid system or else by

decide the feature dependence issue within the naïve bayes classifier, otherwise hybrid the Immune through harsh sets. At long last hybrid systems appear to be the most efficient approach to generate a successful anti spam filter these days.

The future efforts would be extended towards: Achieving accurate classification, with zero percent (0%) misclassification of Ham E-mail as Spam and Spam E-mail as Ham. The efforts would be applied to square Phishing E-sends, which carries the phishing attacks and now-days which is more matter of concern. Also, the work can be extended to keep away the Denial of Service attack (DoS) which has now, emerged in Distributed design called as Distributed Denial of Service Attack (DDoS).

REFERENCES

[1] Issam dagher, Rima Antoun," Ham- Spam Filtering Using DIFFERENT PCA SCENARIOS", 2016 IEEE International Conference on Computational Science and Engineering, IEEE International Conference on Embedded and Ubiquitous Computing, and International Symposium on Distributed Computing and Applications to Business, Engineering and Science

[2] Sch¨olkopf, B., Smola, A.J.: Learning with Kernels. MIT Press, Cambridge (2002)

[3] Ali, S., Smith-Miles, K.A.: A meta-learning approach to automatic kernel selection for support vector machines. Neurocomputing 20(1-3), 173–186 (2006).

[4] Spam (electronic), http://en.wikipedia.org/wiki/Spam_%28electronic%29 Vapnik, V.: Statistical Learning Theory. John Wiley and Sons (1998).

[5] 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.

[6] S Gupta, M Baghel, A Review of Number Plate Recognition Using Neural Network and Feature Extraction Based Technique, EUSRM, 6(1), 2014.

[7] S. Whittaker, V. Bellotti and P. Moody, "Introduction to this special issue on revisiting and reinventing e-mail", Human-Computer Interaction, 20(1), 1-9, 2005.

[8] E-mail spam, http://en.wikipedia.org/wiki/E-mail_spam

[9] 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.

[10] 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.

[11] Yoo, 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.

[12] Mishne, G., Carmel, D., Lempel, R., Blocking blog spam with language model disagreement. In Proc. 1st AIRWeb, Chiba, Japan.

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

[14] Steve Webb, James Caverlee, Calton Pu, 2006. Introducing the Webb Spam Corpus: using Email spam to identify web spam automatically, CEAS.

[15] Sculley, D., Gabriel M. Wachman, 2007. Relaxed online VSMs for spam filtering, SIGIR 2007 Proceedings.

[16] Bing Zhou, Yiyu Yao, Jigang Luo, 2010. A three-way decision approach to email spam filtering. Canadian Conference on AI, pp. 28–39.

[17] Mengjun Xie, Heng Yin, Haining Wang, 2006. An effective defense against email spam laundering, CCS'06, October 30–November 3,Alexandria, Virginia, USA.

[18] Cagri Ozcaglar, 2008. Classification of email messages into topics using latent dirichlet allocation, Master thesis, Rensselaer Polytechnic Institute Troy, New York.

[19] Carmona-Cejudo, Jose´ M., Baena-Garcı´a, Manuel, Morales Bueno, Rafael, Gama, Joa˜ o, Bifet, Albert, 2011. Using GNUsmail to compare data stream mining methods for on-line email classification. J. Mach. Learn. Res. Proc. Track 17, 12–18.

[20]. Bhagyashri U. Gaikwad and P. P. Halkarnikar, " Spam E-mail Detection by Random Forests Algorithm", ISSN (Print): 2278-5140, Volume-2, Issue – 4, 2013.

[21]. Niclas Englesson, "Logistic Regression for Spam Filtering", Mathematical Statistics Stockholm University Bachelor Thesis 2016:9 http://www.math.su.se.