ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

The Upshot of Malicious and Non-Malicious Mails Using Weka

A Sesha Rao, Prof. P.S.Avadhani, D. Chandrika

Research Scholar, Dept. of Computer Science and Systems Engineering, A .U. College of Engineering, Andhra University, Visakhapatnam, Andhra Pradesh, India.

Professor & Principal, College of Engineering, Andhra University, Visakhapatnam, Andhra Pradesh, India.

Asst. Prof, Dept. of Computer Science & Engineering, Vignan's Institute of Engineering for Women, Visakhapatnam, Andhra Pradesh, India.

ABSTRACT: Over the last decades, internet became the primary source for transporting spam and malicious e-mails for numerous criminal activities. So, the email data have become an important target of computer forensics from which we can mine out and analyze the usable clues. Spam mails are invading users without their consent and filling their mailboxes daily. They chomp more network capacity as well as time in checking and deleting spam mails. Data mining classification algorithms are used to categorize the email as spam or hams. Experiments in WEKA environment are conducted using three Classifying algorithms namely ZeroR, J48 and Decision Tree on the spam email dataset and later the three algorithms are compared based on their Evaluation Criteria. In this paper we upshot the email data using ZeroR algorithm, Decision Tree and J48 Cross Validation techniques. We employed supervised machine learning techniques to sift the email spam messages. Our Evaluation Criteria shows noteworthy performance in terms of classification accuracy.

I. INTRODUCTION

Electronic Mail (E-Mail) has emerged as the most important application on internet for communication of messages, delivery of documents and carrying out for transactions. However, security loopholes in it enable cyber criminals to misuse like by forging its headers or by sending secretly for illegal purposes, etc. leading to e-mail forgeries. In effect an overwhelming amount of spam is flowing into user's mail boxes daily. In order to address the growing problem, each organization must analyze the tools available to determine how best to counter spam in its environment. In general, the sender of a spam message pursues one of the following tasks: to advertise some goods, services of ideas, to cheat users out of their private information, to deliver malicious software, or to cause a temporary crash of a mail server. To solve the spam problem, there have been several attempts to detect and filter the spam email on the client side. The information and analysis of E-mail statistics 2015-2019 – Executive Summary report, based on primary research was conducted by The Radicati Group, Inc. It reported that over next four years, the average number of email accounts per user ratio will grow from an average of 1.7 to 1.9 accounts. In the year 2015, the number of emails sent and received per day is around 205 billion [1]. In this paper we upshots the email data using ZeroR algorithm, Decision Tree Table and J48 Cross Validation techniques.

A. About Weka Tool

WEKA is a comprehensive tool bench for machine learning and data mining. Its main strength lies in the classification area. It is developed at the University of Waikato, New Zealand. Weka supports many different standard data mining tasks such as data preprocessing, classification, clustering, regression, Visualization, and feature selection.

Weka is an open source application that is freely available under the GNU general public license agreement. Weka consists of four graphical user interface modules available to the user. They are called Explorer, Experimenter, Knowledge Flow, and Simple Command Line interface. There are some constraints in Weka as it does not accept data in every format. Weka expects data file to be in Attribute-Relation File Format (.ARFF) file. Before we apply the algorithm to our data we need convert data into Comma-Separated file into ARFF file format.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

B. Data sets in WEKA

WEKA accepts the data in Attribute-Relation File Format (ARFF), Comma Separated Value (CSV) as mentioned. Though, it can accept data in CSV format also and can be converted into ARFF format. ARFF file consists of @ RELATION <relation.name> gives the relation declaration, @ ATTRIBUTE <attribute.name> depicts attribute declaration showing the name of attribute and its data type, @DATA illustrates the data declaration that is the start of the data segment in the file. In WEKA, there is option for importing the data as well as generating it automatically [2]. Assume that data stored in Microsoft Excel spread sheet, 'Spam-Data.arff'. After conversion the ARFF file contents look like.

SPAM E-MAIL DATABASE
NUMBER OF INSTANCES: 4601 (1813 SPAM = 39.4%)
NUMBER OF ATTRIBUTES: 58 (57 CONTINUOUS, 1 NOMINAL CLASS LABEL)
CLASS DISTRIBUTION:
SPAM 1813 (39.4%)
NON-SPAM 2788 (60.6%)
1, 0. | SPAM, NON-SPAM CLASSES
WORD_FREQ_MAKE: CONTINUOUS.
WORD_FREQ_ADDRESS: CONTINUOUS.
WORD_FREQ_MAIL: CONTINUOUS.
WORD_FREQ_RECEIVE: CONTINUOUS.
WORD_FREQ_EMAIL: CONTINUOUS.
WORD_FREQ_YOU: CONTINUOUS.
WORD_FREQ_CREDIT: CONTINUOUS
……...

II. LITERATURE SURVEY

Classification based algorithms generally use learning algorithms. A number of spam filtering solutions are proposed in the recent past years. Though the problem of email filtering is not a new one and there exist some implementations to the problem, it is worth highlighting that those implementations have various tradeoffs, difference performance metrics, different classification methods and different performane efficiencies.

Authors Walaa Gad & Sherine Rady [3] has proposed supervised classification and mutual information for feature selection for efficiently classifying spam and ham mails. A content-based approach is presented using Term Frequency (TF) to represent and index email documents.

In [4] the authors investigated feature selection as a preprocessing step using different methods like Chi-Square, Information gain, Gain ratio, Symmetrical uncertainty, Relief, One R and Correlation

Methods used for performance of spam classification by authors [5] are based on DIA and NGL coefficient and compared the results. DIA is an indexing approach which was extended and applied in text categorization. NGL coefficient is an improvement of Chi-Square method. Their experiments show that optimal classification accuracy of 98.5% was achieved at feature length of 104 when model was built using Random Forest classifier.

The authors Tich Phuoc Tran et el [6]; proposed a method for anti-spam filtering problem by combining a simple Linear Regression (LR) model with a Modified Probabilistic Neural Network (MPNN) in an adjustable way. This frame work takes the advantage of the virtues of both the modles. LR-MPNN is shown empirically to achieve better performance than other conventional methods for most of cost-sensitive scenarios.

D.Pliniske's [7] in his research implemented the neural network mehod to the classification of spam mails. His method employs attributes consists of descriptive characteristics of the evasive patterns that spammers adopt rather than using the context or frequency of key words in the message.

Rachana Mishra, R.S.Thakur, [8] analyzed different data mining tools such as WEKA, Rapid Miner and Support Vector Machine. This paper recommends WEKA tool for spam filtering and WEKA outperforms the other data mining tools.

Sujeet More and G.RaviKalkundri [9], the authors used WEKA interface in their integrated classification model and tested with different classification algorithms. They have tested 7 different algorithms such as: Naïve Bayer, Neural Network, Random Forest, Decision Tree, SVM, etc. They found Random Forest & SVM classifiers outperforms the conventional one.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

Recent years email spam filtering made considerable progress. Many solutions have been proposed for spam mail identification and in which the White-list and Black-ist [10] filtering methods are operated based on IP Address, DNS, or Email Address. These filtering methods maintain the source of received spam in a data base with priority. Each time when a new mail received its source is compared with the contents of data base. The main disadvantage is when spamers regularly change email and IP addresses to cover their trails. [11,12] shows different machine learning methods for spam detection.

III. PROPOSED WORK

In this paper with the support of WEKA TOOL we implemented the proposed three algorithms for identifying legal and ill legal data.

To classify with WEKA Graphic User Interface (GUI) the basic steps are:

1. Run WEKA GUI
2. Click 'Explorer/Knowledge Flow'
3. 'Open file,'
4. Select 'Classify Tab'
5. Choose a 'Classifier'
6. Confirm options
7. Click 'start'
8. Right Click on, Result list entry
   a) Save result buffer
   b) Save model

The architecture of the proposed system is shown below. The architecture describes opening WEKA tool and selecting Knowledge Flow option followed by data source and Data sink fields to convert files to WEKA acceptable format.

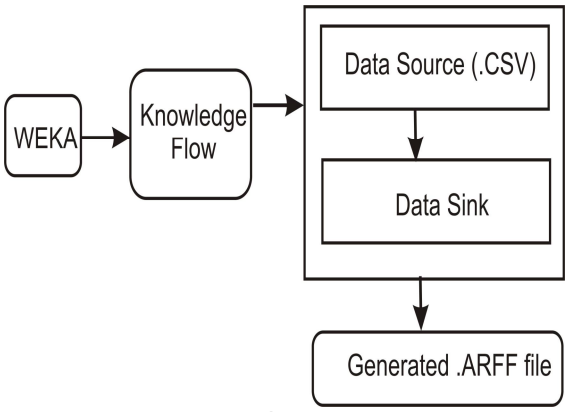

Fig. 1 WEKA Preprocessing

Knowledge Flow: Knowledge Flow is a Java-Beans-based interface for setting up and running machine learning experiments [2]. With the KnowledgeFlow interface, users select WEKA components from a tool bar, place them on a layout canvas, and connect them into a directed graph that processes and analyzes data. It allows the design and execution of configurations for streamed data processing, which the Explorer cannot do.

Data Source: This takes the source file (.CSV file) as input.

In order to experiment with the application the data set needs to be presented to WEKA in a format that the program understands. There are three options for presenting data into the program.

♦ Open File- allows for the user to select files residing on the local machine or recorded medium
♦ Open URL- provides a mechanism to locate a file or data source from a different location specified by the user
♦ Open Database- allows the user to retrieve files or data from a database source provided by the user

The supported data formats are .ARFF, .CSV, C4.5 and binary. Alternatively you could also import data from URL or an SQL database. After loading the data, pre-processing filters could be used for adding/removing attributes, discretization, Sampling, randomizing etc.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

Pre-processing *is* used to choose the data file to be used by the application

Data Sink: This is the destination format to which we want to convert into .ARFF file format
WEKA expects the data file to be in Attribute-Relation File Format (ARFF) file. In fact once loaded into WEKA, the dataset can be saved into ARFF format.
Generate ARFF file: Enables to generate artificial data from a variety of Data Generators.

IV. UPLOAD .CSV FILE FROM OUTLOOK TO EXCEL SHEET

In order to train and test the proposed system some corpora of spam and legitimate emails had been used by collecting the mails from Microsoft Outlook. MS Outlook is a different application from Outlook Express. The two programs do not share a common codebase, but do share a common architectural philosophy.
Microsoft Outlook uses a proprietary email attachment format called Transport Neutral Encapsulation Format (TNEF) to handle formatting and other features specific to Outlook such as meeting requests [13]
Outlook stores all mail information in a single file with an extension '.pst'. A 'personal storage table' file stores all account data like mails, contacts, journal, notes and calendar entries.
Export contacts from MS Outlook 2007 using the following procedure.

- From the Outlook main menu, select **File** > **Import and Export**. The Outlook "Import and Export" Wizard displays.
- Select **Export to a file** and then click **Next**.
- Select **Comma Separated Values (Windows)** and then click **Next**.
- Choose to export from the **Contacts** folder and then click **Next**.
  Note: If you have your email addresses in an Outlook Personal Address Book, first convert your email Personal Address Book to a Contacts folder. See your Outlook online help for more information.
- Type a file name (ex: MycontactList) under "Save exported file as" and click **Browse** to locate the directory where you want to place the exported file. Then click **OK** to close the "Browse" dialog box.
- Click **Next**.
- Click **Map Custom Fields...**
- Use the columns to map values from your Microsoft Office information to the file you are exporting.
- Click **OK**.
- Click **Finish**. The new .CSV file should now be in the location that you specified.

- **Note:** Prepare your list. If you only want email addresses and contact names, remember to open the file in MS Excel and strip the list of any extra information that may have been exported.
- Save the file into Constant Contact.

Now the .CSV file is generated for the e mail data in Excel Sheet which will be used for preprocessing in WEKA.
Email preprocessing is the process of transforming email messages into a suitable representation suitable for the preparation of training datasets. The preprocessing involves tokenization, noise symbols elimination, stop words elimination, suffix stripping etc.

V. ARCHITECTURE FORUPSHOT MAIL CLASSIFICATION

Classification is a machine learning technique used to predict group membership for data instances. [14] It is the prediction of finding the model for class attribute as a function of the values of other attributes and predicting class assignment for test data. Classification is a two-step process: first is model construction i.e. describing a set of predetermined classes and second is using that model for prediction i.e. classifying future or unknown instances.
The architecture of our proposed model is shown in figure.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

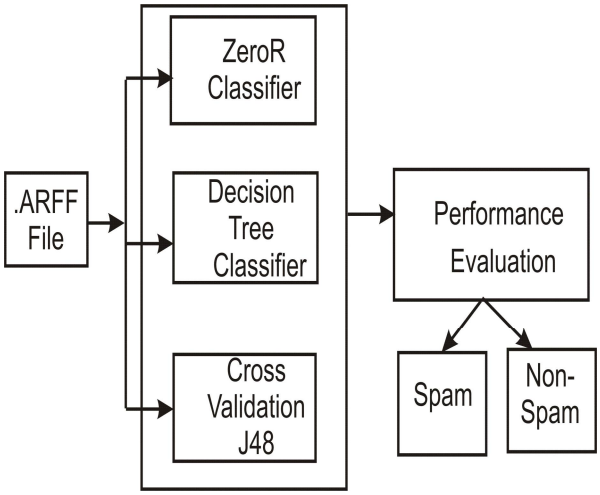

Fig. 2 Architecture for Upshot mail classification

Some of the classification techniques that are supported by WEKA tool are discussed below:
a). Basic ZeroR Classifier: ZeroR is a simplest classification method which relies on the target ignores all predictors. It Predicts the mean (for a numeric Class) or the mode (for a nominal Class). It simply predicts the majority category. It constructs a frequency table for the target and selects its most frequent value. There is nothing to be said about the predictor's contribution to the model because ZeroR does not use any of them.

b). J48 Classifier: J48 builds decision tree from a set of training data using the concept of information entropy. [15] J48 examines the normalized information gain. It uses the fact that each attribute of the data can be used to make a decision by splitting the data into smaller subsets. J48 classifier recursively classifies until each leaf node is pure, which means the data has been perfectly classified. It choses one of the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The attribute with the highest normalized information gain is chosen to make the decision. The algorithm then returns on the smaller subset.
This algorithm has a few base cases-
A leaf node is created when all the samples in training set belong to the same class. Then the decision tree choses that class.
If the features does not provide any information gain, then J48 creates a decision node higher up the tree with the expected value of class.
Also J48 creates a decision node higher up the tree with the expected value, when previously unseen instance class is encountered.

c).Decision Tree: A decision tree is a graphical model describing decisions and their possible outcomes.
A decision tree is classification method that results in a flow chart like tree structure where each node denotes a test on attribute value and each branch represents an outcome of the test. The tree leaves represents the classes [16].

Decision Tree consists of three types of nodes.
1. Decision node: Often represented by squares showing decisions that can be made. Lines emanating from a square show all distinct options available at a node.
2. Chance node: Often represented by circles showing chance outcomes. Chance outcomes are events that can occur but are outside the ability of the decision maker to control.
3. Terminal node: Often represented by triangles or by lines having no further decision nodes or chance nodes. Terminal nodes depict the final outcomes of the decision making process.
Decision tree use divide and conquer method to split the problem search space into subsets. There are two basic steps in this technique: building the tree and applying the tree to dataset.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

Cross Validation: There are several means of estimating how well the classifier works after training [15]. The trained classifier can be evaluated either with an additional test set or through k-fold cross validation, or by dividing the input dataset to a training and test set. In the k- fold cross validation, the data set M is portioned into k mutually exclusive parts, $M_1,M_2,\ldots.M_k$. The inducer is trained on Mi /M and tested against Mi. This process is repeated k times with different values of i. Finally the performance is estimated as the mean of the total number of tests. For a k fold test the precision *P* and the recall *R* are defined as

$$P = \frac{1}{n} \sum_{i=1}^{k} p_i$$

$$R = \frac{1}{n} \sum_{i=1}^{k} r_i$$

Where p_i and r_i are the precision and recall for each of the k tests. The research has shown that k=10 are a satisfactory total, therefore 10 fold cross validation is was used throughout the experiments.

VI. EXPERIMENTAL EVALUATION

In this section we compare the classification accuracy results of the three proposed classification algorithms, namely, ZeroR, Decision Tree Tables and J48. The simulations are conducted using a large spam e-mail dataset collected from MS Outlook as mentioned above. The dataset consists of a total of 4601 mails, from which 1813 are spam emails and 2412 are ham emails. The dataset has been divided into training and test datasets where a 10 cross fold was used. [3] The data set is randomly portioned into 10 equal-sized folds or subsets. Then the classifier is trained on 9 folds and the remaining fold is used for testing. The cross-validation process is then repeated until each of the 10 folds is used once in the testing fold.

Load the dataset 'Spam-data.arff ' into WEKA, perform a series of operations using WEKA's attribute and discretization filters. This is shown in Fig. a. after loading dataset.

Choosing Classifier: Once the dataset is loaded all the tabs are available to you, Click on the Classify' tab, then "Classify" Window come up on the screen. Now you can start analyzing the data using the provided algorithms. We analyzed the data with ZeroR, Decision Tree and J48. The sample data used in this exercise is "Spam-data.arff". Click on 'Choose' button in the classifier box just below the tabs and select C4.5 classifier then follow the path:

WEKA → Classifier → Trees → J48

Now you need to set test options. Set 'test options' in the 'test options box'. The test options that available are training set, supplied test set, cross-validation and percentage split.

Evaluating Spam and non-Spam data using basic ZeroR algorithm which is shown in Fig. 3
Cross Validation is a Built-in automatic method of self-testing a model for reliability using Decision Tree. It is to reduce variance where we can improve upon repeated holdout.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

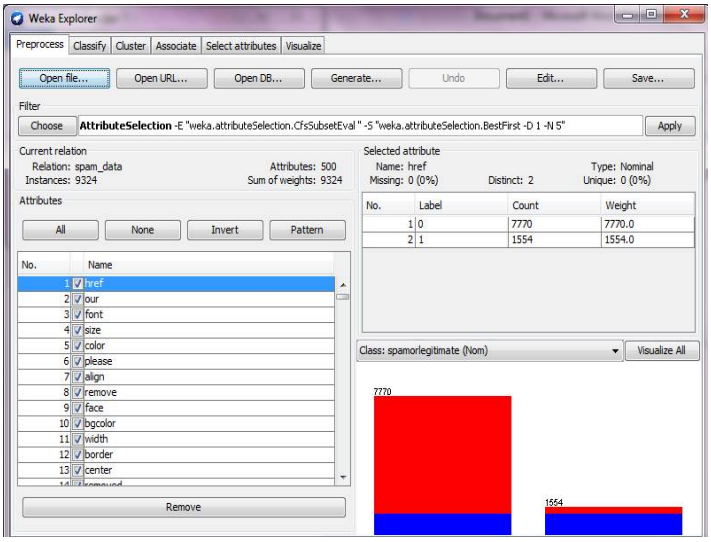

Fig. 3 Screen-1: Loading the Data file Into WEKA Tool

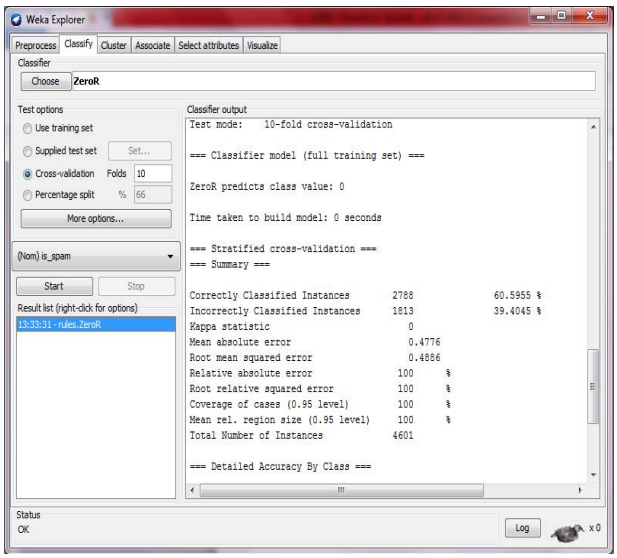

Fig. 4 Screen-2:WEKA Outputs for ZeroR Algorithm

.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

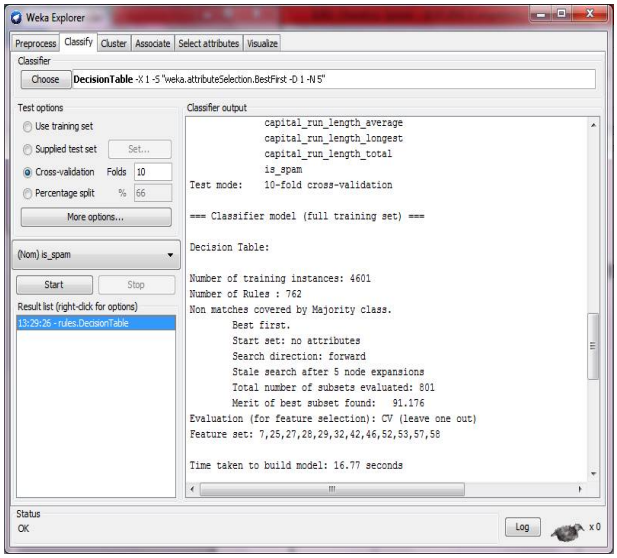

Fig. 5 Screen-3 WEKA Outputs of J48 Algorithm.

VII. PERFORMANCE EVALUATION

To evaluate the classifier on testing spam data we define the following measures are have been used; True Positive (TP Rate), False Positive (FP Rate), False Negative (FN Rate) Precision, Recall, Prediction, accuracy, Execution time, F-Measure, MCC, ROC Area, PRC Area. The performance measures are defined as:

Precision: It is the fraction of retrieved instances that are relevant.

$$\text{Precision}: \frac{TP}{(TP+FP)}$$

Recall: It is the fraction of relevant instances that are retrieved.

$$\text{Recall}: \frac{TP}{(TP+FN)}$$

Accuracy: Accuracy refers to the closeness of a measured value to a standard or known value.

$$Accuracy: \frac{(TP+TN)}{(TP+FP+TN+FN)}$$

Where TP is True Positive that represents the classifier identifying a positive class data instance as positive; FP is False Positive that represents the classifier has identified a negative class data instance as positive; FN is False Negative that represents the classifier has identified a positive class data instance as negative; TN is True Negative that represents the classifier has identified a negative class data instance has negative. The accuracy is the proportion of true results (both TPs and TNs) among the total number of cases examined.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

F-Measure (Frequency Measure or F-Score): It is the harmonic mean of its precision and recall.

$$\text{F-score} = \frac{2 \times Precision \times Recall}{(Precision + Recall)}$$

Matthews Correlation Coefficient(MCC):

MCC used in machine learning as a measure of the quality of binary (two-class) classifications, introduced by biochemist BrianW.Matthewsin. A coefficient of +1 represents a perfect prediction, 0 no better than random prediction and −1 indicates total disagreement between prediction and observation. The MCC is in essence a correlation coefficient between the observed and predicted binary classifications; it returns a value between −1 and +1. It takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

Receiver Operating Characteristic (ROC): In statistics, it is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied.

Precision Recall Curve(PRC): Precision-recall curves are typically used in binary classification to study the output of a classifier.

Table 1 lists the performance results for the different classifiers, in terms of TPR, FPR, precision, recall, ROC area etc. Parameters of average weight are True Positive Rate, False Positive Rate, Precision, Recall, F- measure, Mathew's Correlation Coefficient, ROC Area, and PRC Area. Values ranging between 0 and 1 are obtained indicating the intensity of spam. 0 indicates that the presence of spam is nil while 1 indicates that the presence of spam is high. The following are the experimental results that are obtained when computed using WEKA tool.

Table1: Performance Results of three algorithms.

Classifier	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
ZeroR	0.606	0.606	0.367	0.606	0.457	0	0.499	0.522
Decision Tree	0.9	0.125	0.904	0.903	0.902	0.796	0.948	0.949
J48	0.93	0.078	0.93	0.93	0.932	0.853	0.939	0.917

The following is a column graph representing the existence of spam over the three algoithms – ZeroR, Decision Tree and J48.

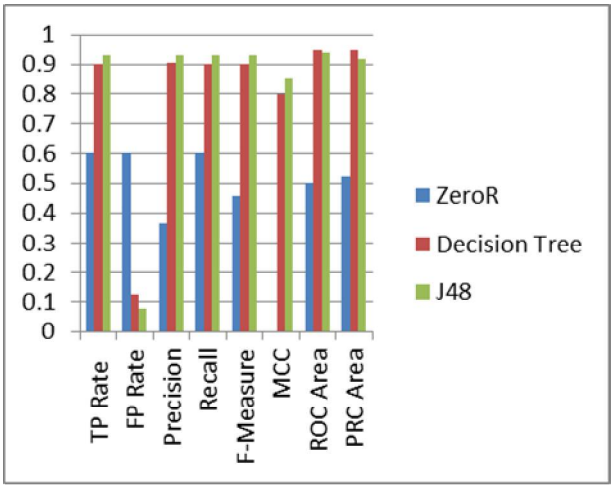

Fig. 6 Review results of the ZeroR, Decision Tree and J48 on the Spam Dataset .

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

Evaluation of classifiers using 10 Folds Cross Validations on parameters like Time taken to build the classifying model, Correctly Classified Instances, & Predication Accuracy, Incorrectly Classified Instances & Predication Accuracy are tabulated in the following table.

Table 2: Final Statistics of the three Algorithms Using Cross Validation (10-Folds)

Evaluation Criteria	Classifiers		
	ZeroR	Decision Tree	J48
Time taken to build the classifying model	0.02 Sec	13.81Sec	3.64Sec
Correctly Classified Instances & Predication Accuracy	2788, 60.59 %	4155, 90.31 %	4278, 92.98 %
Incorrectly Classified Instances & Predication Accuracy	1813, 39.40%	446, 9.69 %	323, a.

VIII. CONCLUSION

Spam is big problem of today's world, to solve the problem several research papers have been evolved. In this paper we upshot the email data using ZeroR algorithm, Decision Tree and J48 Cross Validation techniques. We used the platform WEKA tool to implement the proposed classification algorithms to detect spam or ham mails. The proposed work will be helpful for identifying the deceptive mail and will also assist the investigators of email forensics. The Decision Tree and J48 give better performance. The major advantage of the Decision-Tree based classifier is that it doesn't assume that terms are independent and its training is relatively fast,

IX. FUTURE WORK

In our future work we want to adopt Artificial Neural Network approach with Krill Herd algorithm which is an improved optimization technique.

REFERENCES

[1] http://www.radicati.com
[2] WEKA User Manual; http//www.gtbit.org/downloads/dwdmasem6/dwdmsem6lman.pdf
[3] Walla Gad and Sherine Rady "Email Filtering based on Supervised Learning and Mutual Information Feature Selection", IEEE Conference, 2015, 978-1-4673-9971-5/15.
[4] P. Ozarkar and Dr. M. Patwardhan,"Efficient Spam Classification By Appropriate Feature Selection", International Journal of Computer Engineering and Technology (IJCET), ISSN 0976 –6375 (Online) vol.4(3), May – June 2013.
[5] Josin Thomas, Nisha S. Raj, Vinod P., *"Robust Feature Vector for Spam Classification"*, In proceedings of the International Conference on Data Sciences, Universities Press, ISBN:978-81-7371-926-4, February 2014, pp. 87-95
[6] Tich Phuoc Tran, Pohsiang Tsai, Tony Jan, "An Adjustable Combination of Linear Regression and Modified Probabilistic Neural Network for Anti-Spam Filtering" IEEE Conference 2008, 978-1-4244-2175-6/08.
[7] D. Puniškis, R. Laurutis, R. Dirmeikis, An Artificial Neural Nets for Spam e-mail Recognition, electronics and electrical engineering ISSN 1392 – 1215 2006. Nr. 5(69)
[8] Rachana Mishra, Ramjeevan Singh Thakur, "An Efficient Approach for Supervised Learning Algorithms using Different Data Mining Tools for Spam Categorization", IEEE, 2014, 978-1-4799-3070-8/14
[9] Sujeet More, Ravi Kalkundri, "Evaluation of Deceptive Mails Using Filtering & Weka", IEEE 2015, 978-1-4799-6818-3/15
[10] Spam Cop, Spam Cop Blocking List.Available:http://www.spamcop.net/bl.shtml, 2010.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

[11] DeBarr, H.W.D., Spam Detection using Clustering, Random Forests and Active Learning, presented at the 6th Conference on Email and Anti-Spam, California, 2009
[12] Awad, S.M.E.W.A., Machine Learning methods for Email Classification, International Journal of Computer Applications, 2011.
[13] Christian Bird, Alex Gourley, Prem Devanbu, Michael Gertz and Anand Swaminathan, "Mining Email Social Metworks" (May 22-23m 2006). Dept. of Computer Science, Unviersity of California, Davis.
[14] Sadeghian, A and Ariaeinejad, R. , "Spam detection system: A new approach based on interval type-2 fuzzy sets", iEEE CCECE -000379, 2011.
[15] Ms.D.Karthika Renuka ,Dr.T.Hamsapriya , Mr.M.Raja Chakkaravarthi ,Ms.P. Lakshmi Surya, "Spam Classification based on Supervised Learning using Machine Learning Techniques", IEEE 2011, 978-1-61284-764-1/11.
[16] Megha Rathi, Vikas Pareek, " Spam Mail Detection through Data Mining – A Comparative Performance Analysis" I.J. Modern Education and Computer Science, 2013, 12, 31-39.

BIOGRAPHY

A.Sesha Rao, is a research scholar in the department of Computer Science and Engineering, College of Engineering, Andhra University, Visakhapatnam. His research interests are related but not limited to cryptography, security, privacy and email forensics. In 1972, he received M.Sc. (Applied Maths) from Andhra University, Waltair, M.Tech in Computer Science & Engineering, from IIT, Mumbai, 1985. After that he worked as Scientist for 25 years in Naval Science and Technological Lab, DRDO, Visakhapatnam, presently he is working as Professor in Vignan's Institute of Engineering for Women, Visakhpapatnam.

Dr.P.S. Avadhani is a Principal AU College of Engineering (Autonomous) . He has guided 17 Ph.D. students, 3 students already submitted and right now he is guiding 12 Ph.D. Scholars from various institutes. He has guided more than 100 M.Tech Projects. He has 15 Journal Publications, 62 Conference Publications. He has co-authored 4 books. He received many honors and he has been the member for many expert committees, member of Board of Studies for various Universities, Resource persons for various organizations. He is a Life Member in CSI, AMTI, ISIAM, ISTE, YHAI and in the International Society on Education Technology. He is also a Member of IEEE, and a Member in AICTE.

D.Chandrika is an Assistant Professor of Vignan's Institute of Engineering for Women in the department of Computer Science and Engineering. She is a member of IEEE and IAENG.