A Comparative Performance Evaluation of Content Based Spam and Malicious URL Detection in E-mail

Sunil B. Rathod
PG Student, Department of Computer Engineering,
North Maharashtra University,
SES's R. C. Patel Institute of Technology, Shirpur, India
sunilrathod.rathod01@gmail.com

Tareek M. Pattewar
Assistant Professor, Department of Computer Engineering,
North Maharashtra University,
SES's R. C. Patel Institute of Technology, Shirpur, India
tareekpattewar@gmail.com

***Abstract*—E-mail communication is growing rapidly. Email contains Text and URLs as content. Text can be suspicious, from undesired sender which contains un-required content and URLs may be malicious which redirects users to phishing (malicious) websites. Thus to stop such activity a spam and malicious URLs detection system is required which benefits users by removing spam content and malicious URLs in Email. We have used data mining approach like supervised classification which improves the systems accuracy and detects more amount of spam and malicious URLs.**

Keywords— Bayesian Classifier, Decision Tree, Malicious URL Detection, Spam Detection .

I. INTRODUCTION

EMail is becoming fastest and economical mode of communication . The growing use of email has lead to increased rate of spam emails. As it is information age users rely on emails to communicate with the globe. Business organization, individuals and all corporate industries are communicating with emails so that it is important part concerning with education, business and personal usage.

Spam:

Spam are nothing but the unsolicited bulk emails (UBE) and it's another part is unsolicited commercial email .These spam emails not only consume the user's time but also the energy to recognize the undesired messages, It is wasting the network bandwidth.

Content Based Spam Filter:

Content Based filter works on content of emails i.e., text, URLs, main headers like subject for classification purpose. It is the method used to filter spam.

The emails include two parts such as Body of the message and Header, Header stores the information about message like from whom it is received, date and time of emails received, sender etc. Now emails ambiguous data is removed by preprocessing then text is extracted.

978-1-4673-7437-8/15/$31.00 ©2015 IEEE

II. RELATED WORK

Today's internet is suffering from major problem known as Email spam .It annoys users and make financial damage to companies. So far developed techniques to stop spam are filtering methods .Spam emails are UBE also known as junk emails ,that are send to many recipients who have not requested or subscribe to this. Spam filter removes spam or un-required messages from email inbox . It also has Phishing URLs which redirects users to phishing websites and seeking personal credentials like username and password for financial purpose.

The existing work by Dhanalakshmi R and Chellapan C, did implementation on malicious URL detection in Email. Lexical features, page rank, Host information are taken into consideration to classify URLs. Phishtank corpora has been used and Bayesian classification is done to improve the performance of system [1].

Georgios Paliouras et al., have presented learning method to filter spam email. The two machine learning algorithm are considered for anti-spam filtering such as Naïve Bayesian and Memory based learning approach and they are compared concerning performance. So, that in both methods spam filtering accuracy has improved and keyword based filter are used widely for email [2].

Zhan Chuan, LU Xian-liang has given an application for email filtering using a new improved Bayesian filter. They have represented word frequency by vector weights and word entropy is used for attribute selection then formula is derived which improves the performance apparently [3].

Vikas P. Deshpande et al., has presented an efficient method of naïve Bayesian which blocks all spam emails without blocking legitimate emails. To derive solution on this problem, they considered statistical classifier such as naïve Bayesian anti-spam filter and content based spam filter which are adaptive in nature [4].

Sheng et al., have shown that phishing websites are hacked as soon as they are identified as phishing campaigns have two hours of average life. So to block and identify such phishing URLs they have extracted features like suspicious characters, number of dots, ip address, hexadecimal character [5].

Pawan et al., discovered malicious URLs by enhancing blacklisting. One conflict with this method is that their updation process is fast so they failed to identify phishing URLs in early hours of a phishing attack[6].
Maher Abburrous et al., endeavor for a survey to recognize the essential features which can develop accuracy and precision for malicious URLs detection [7].
Congfu Xu et al, did a feature extraction on Base64 encoding of image with n-gram technique. A SVM needs to be trained for efficiently detecting spam images from legitimate images. Its seen from experiment that It has improved the performance in terms of Accuracy, Precision and Recall [8].
R. Malathi et al., has given a new spam detection method by employing Text Categorization, using Supervised Learning with Bayesian Neural Network which uses Rule based heuristic approach and statistical analysis tests to identify "Spam" [9].
Sadeghian A. et al, had presented spam detection based on interval type-2 fuzzy sets. This system gives user more control on categories of spam and permits the personalization of the spam filter [10].
CANTINA+ classifies phishing URLs and the feature set is more exhaustive and obtained classification accuracy of 92.3%. There exist various related researches and case studies conducted on analyzing the feature set required to reduce the exhaustiveness and time consumption [11].

III. Algorithm Study

A. Bayesian Classifier:

Naïve bayes classifier is statistical classifier famous for Email filtering, Spam emails are identified by classification method. Naïve bayes uses tokens (words) with spam and ham mails for Calculating probability to determine whether a mail is spam or not.
Mathematical Formulation:
Bayesian classifier is based on Naïve Bayes theorem, Naïve Bayes theorem can perform more sophisticated classification methods.
To demonstrate the concept consider following equations [11];
Thus, we can write:
Prior probability of Legitimate mail = Number of legitimate mail / Total number of mail (3.1)
Prior probability of Spam mail = Number of spam mail / Total number of mail (3.2)
Likelihood of X-mail given Legitimate = Number of legitimate mail in the vicinity of X-mails / Total number of legitimate mail. (3.3)
Likelihood of X-mail given Spam = Number of spam mail in the vicinity of X-mails / Total number of spam mail. (3.4)
Posterior probability of X-mail being legitimate = Prior probability of legitimate mail × Likelihood of X-mail given legitimate. (3.5)
Posterior probability of X-mail being spam = Prior probability of spam mail × Likelihood of X-mail given spam. (3.6)
Finally we classify X-mail as spam as its class membership has a largest posterior probability.

B. Decision Tree C4.5:

C4.5 is developed by Ross Quinlan. It is Extension of ID3 and also known as statistical classifier. C4.5 creats decision tree alike ID3 as it is successor of ID3 using the concept of "Information Entropy": It is measure of homogeneity of a learning set. At each node of tree, C 4.5 selects attributes for dividing its sets into subsets. Normalized information gain is the important criterion for splitting the data. Another term is "Information Gain" which is the difference in information entropy associated with attribute. The attribute with highest normalized information gain is choosen to make decision. performance of the system can be derived by Accuracy and Error Rate as follows;

$$\text{Accuracy} = \frac{\text{No of correctly classified samples}}{\text{Total no. of samples in the class}} \quad (3.7)$$

$$\text{Error} = \frac{\text{No of Incorrectly classified samples}}{\text{Total no. of samples in the class}} \quad (3.8)$$

IV. Experiment

A. Implementation using Bayesian Classifier :

1) Gmail Dataset and SpamAssassin Dataset:

This is the combination of the real time dataset downloaded from Gmail and some emails from SpamAssassin in bulk consisting of legitimate and spam emails. These emails are considered for input to preprocess in HTML format.

2) Text Preprocessing:

a) HTML Tag Removal:

The input Emails are in HTML format so this contains the tag, so to purify the text we need to remove the tags.

b) Stopword Removal:

This is the stopword list which consist of terms including articles, prepositions, conjunctions and certain high frequency words (such as some verbs, adverbs)

c) Tokenization :

Lexical analysis also named as Tokenization, It

involves dividing the content of text into strings of character called as Tokens. Filtering techniques uses white space (blank) removal and removal of punctuation symbols in tokenizing.

d) Word Frequency:

This counts the frequency of words depending on its occurrence, This helps in deriving the word probability for spam and legitimate mails. Or

Term Frequency

Terms Frequency of term can be defined as the overall frequency of a term in the entire corpus i.e. in the entire email instances. To calculate the TF score, frequencies of terms in individual emails were first calculated and then all the frequencies of a term in the entire set of emails were added to find the TF Score for a particular term tk. Mathematically it can be expressed as

$$TF(tk) = \sum_{f=1}^{N} (tf_{k,f}) \qquad (4.1)$$

Terms having less TF Score will be eliminated and those having high score will be selected.

3) Bayesian Classifier:

It is method used for classification of text, It gives efficient learning algorithm for data mining. This uses Bayes classifier theorem which is based on conditional independence assumption:

P (spam/word) = [P (word/spam) P (spam)] / p (word)

Considering spam probability for words, It evaluates Spam and Legitimate mails for classification then gives performance measurement.

4) Performance Measurement

Performance can be evaluated in terms of Accuracy, Error, Time, Precision and Recall for Base method using Bayesian Classifier .

B. Implementation using combination of both Bayesian Classifier and Decision Tree C4.5 :

As Email body consist mainly of 'TEXT' and 'URLs', for TEXT we do classification based on Bayesian Classifier and The process undergo classification as in A) Base Method using Bayesian Classifier and for URLs we use following method of classification.

1) Phishtank Dataset and DMOZ Dataset

Phishtank is source of blacklisted phishing URLs which admits user input and they are verified by users. It is set of URLs which are suspected and reported as phishing URLs to phishtank.

DMOZ: It is used to get genuine and legitimate URLs of web links for Dataset of legitimate or non-phishing URLs.

2) URL Preprocessing :

a) IP Address:

IP addresses and hexadecimal characters are used to hide the actual URLs. For example consider the URL http://www.bankingcompany.com/online/transaction/website/ phishing.html" which is shortened using the IP address http://132.115.201.115 which looks like legitimate and not suspicious.

b) Hexadecimal Character

The URL can also be represented using hexadecimal base values with a '%' symbol. It may represent any special characters Spoofguard identified the '@' and '-' symbol most prominent in phishing URLs. In URL a @ symbol is considered as centre and its left side is dispensed and its right side is thrown into phishing site. Consider the URL http://www.citibank.com@phishingsite.com" will enter into "phishing site.com" and discards "www.citibank.com". Such types of methods uses mask for phishing site and pretense as legitimate sites.

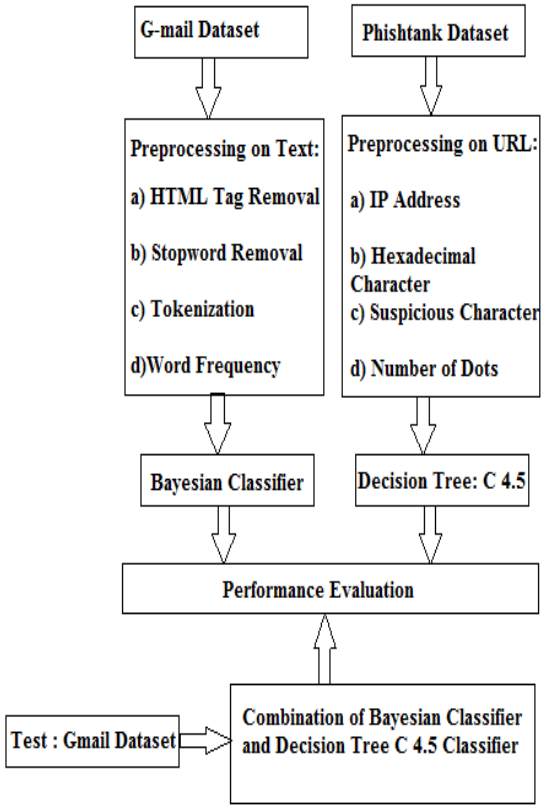

Fig. 1. Combination Approach of Content Based Spam Detection using Bayesian Classifier and malicious URLs Detection in Email using Decision Tree C4.5.

c) Suspicious Character

Presence of suspicious characters such as @ symbol and other special binary characters such as ('.', '=', '$', '^' and etc.) either in the host or path name, can be suspicious characters.

d) Number of Dots

In this number of dots are observed in given URLs of email to predict whether a given URL is malicious or legitimate.

3) Decision Tree C4.5:

The Dataset from Phishtank is preprocessed and passed as input to Decision Tree C4.5 for classification then performance is measured in terms of Accuracy, Time and Error.

4) Testing G-mail Dataset :

This is derived from g-mail consisting of spam and legitimate mails .It also needs to be preprocessed in two terms : A) Preprocessing for Text and B) Preprocessing on URLs to give pure Text and URLs then classification is done by combination of both Bayesian classifier and Decision Tree (C 4.5). Further correctly classified instances (mails) and Incorrectly classified instances (mails) are evaluated.

5) Performance Measurement:

As combination classification model builds of Bayesian and Decision Tree C4.5, It is essential to derive performance on the basis of parameters such as Accuracy (Correctly classified instances), Error (Incorrectly classified instances), precision and Recall are evaluated .

Accuracy = (TN + TP) / (TN + TP + FN + FP)

Error = 100- (Accuracy)

Precision = (TP) / (TP+FP)

Recall = (TP) / (TP + FN)

Where,
TN: True Negative, Legitimate predicted as Legitimate
TP: True Positive, Spam predicted as Spam
FP: Legitimate predicted as Spam
FN: Spam predicted as Legitimate.

V. Experimental Results and Performance Evaluation

A. Computation of system's efficiency under different volume of Dataset for combination approach using Bayesian Classifier and Decision Tree (C4.5) Classifier:

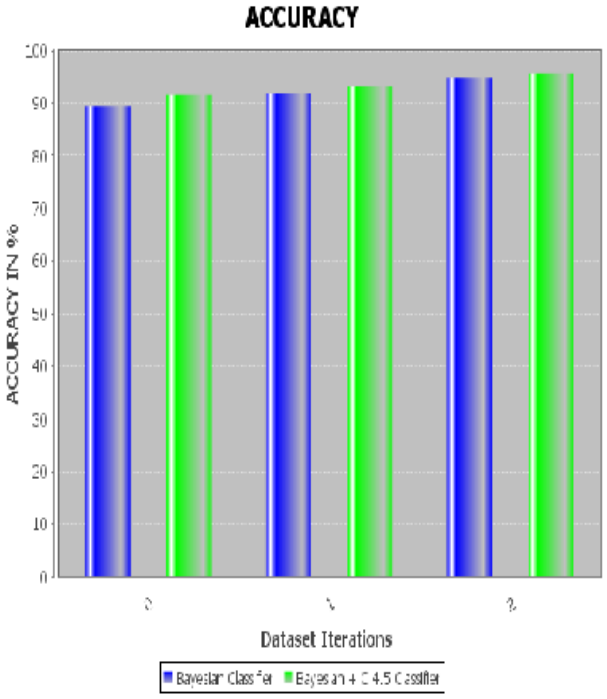

Fig. 2. Accuracy of the Implementation for different volume of the Datasets

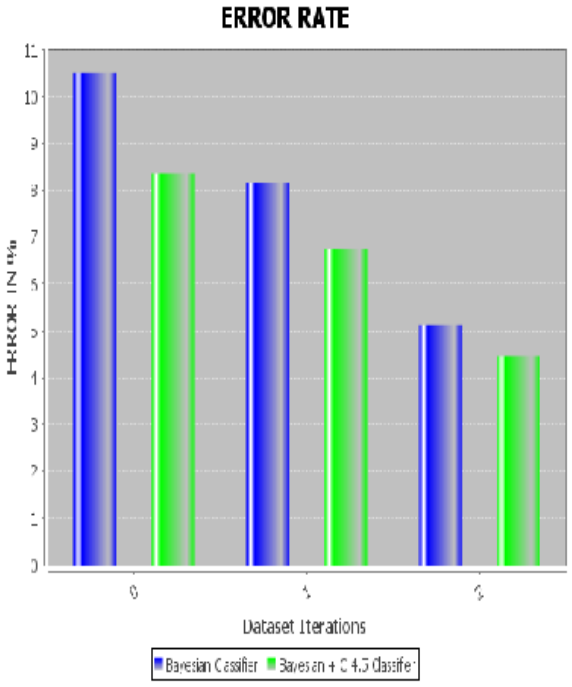

Fig. 3. Error of the Implementation for different volume of the Datasets

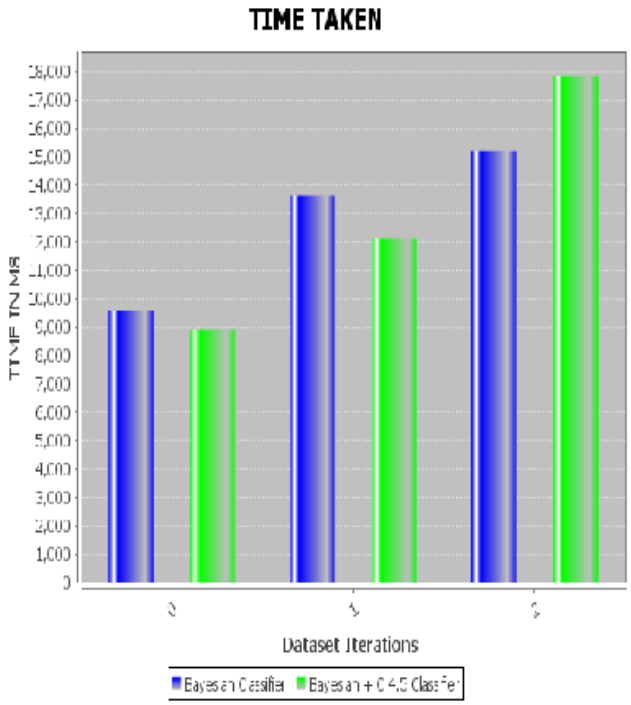

Fig. 4. Time taken for Implementation for different volume of the Datasets

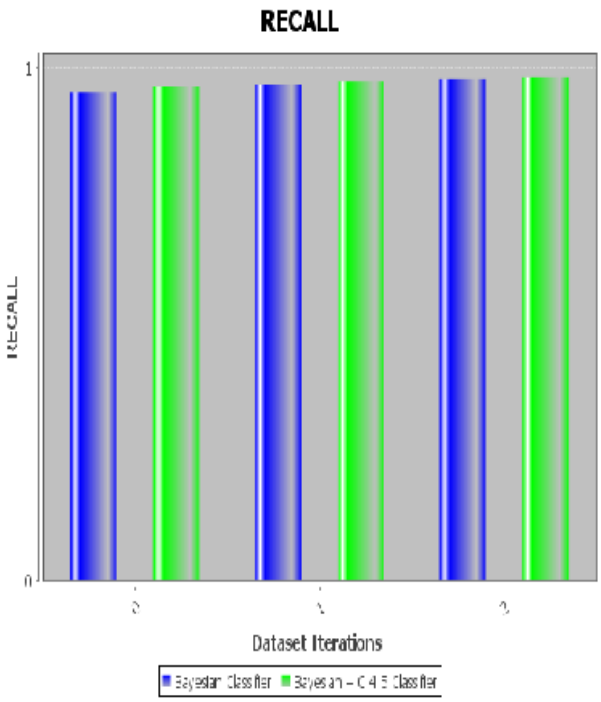

Fig. 6. Recall of the Implementation for different volume of the Datasets

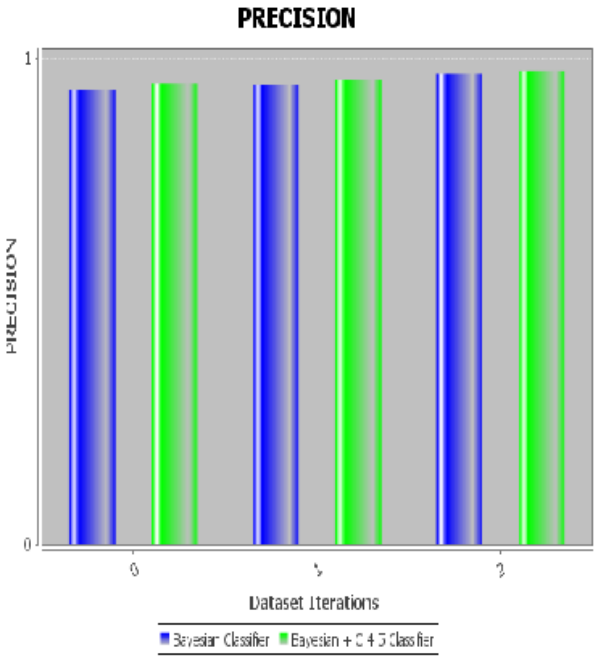

Fig. 5. Precision of Implementation for different volume of the Datasets

B. Tabular Results:

TABLE I

Implementation Results using Bayesian Classifier

Bayesian Classifier	Accuracy (%)	Error (%)	Time (MS)	Precision	Recall
Dataset 1	89.5	10.5	7690.0	0.93	0.95
Dataset 2	91.83	8.17	9563.0	0.94	0.97
Dataset 3	94.86	5.13	22292.0	0.97	0.98

TABLE II

Implementation Results using Combination of Bayesian Classifier and Decision Tree C4.5 Classifier

Bayesian + C 4.5 Classifier	Accuracy (%)	Error (%)	Time (MS)	Precision	Recall
Dataset 1	91.63	8.37	5616.0	0.95	0.96
Dataset 2	93.26	6.74	14243.0	0.95	0.97
Dataset 3	95.54	4.46	25725.0	0.97	0.98

TABLE III
Comparative Performance Evaluation of A) Implementation using Bayesian Classifier and B) Implementation using Bayesian Classifier and Decision Tree (C4.5) Classifier Where, A - Bayesian Classifier and B - Bayesian and C4.5 Classifier

Performance Evaluation	Accuracy (%)		Error (%)		Time (MS)		Precision		Recall	
Classifiers	A	B	A	B	A	B	A	B	A	B
DATASET 1	89.5	91.63	10.5	8.37	7690.0	5616.0	0.93	0.95	0.95	0.96
DATASET 2	91.83	93.26	8.17	6.74	9563.0	14243.0	0.94	0.95	0.97	0.97
DATASET 3	94.86	95.54	5.13	4.46	22292.0	25725.0	0.97	0.97	0.98	0.98

VI. CONCLUSIONS

We have integrated the content based spam detection using Bayesian Classifier and phishing URLs detection using Decision Tree C4.5. Thus we found that performance evaluated for combination approach of Bayesian classifier and Decision Tree C4.5 are improved as compared to implementation using content based spam detection by Bayesian Classifier.

We have evaluated the results across different volume of dataset, Implementation using Bayesian classifier gives 94.86 % accuracy whereas The Combination approach of Bayesian Classifier and Decision Tree C4.5 gives 95.54 % accuracy So, We can say that combination approach has improved the results in terms of Accuracy and It became the efficient method for classification of content based spam detection and malicious URL detection in integrated form.

ACKNOWLEDGMENT

We are sincerely grateful to all the persons who help us through this work to make it successful.

REFERENCES

[1] Dhanalakshmi Ranganayakulu and Chellappan C., "Detecting malicious URLs in E-Mail - An implementation", in *AASRI Conference on Intelligent Systems and Control*, Vol. 4 , pg. 125–131, 2013.

[2] G. Paliouras *et al*.,"An Evaluation of Naive Bayesian Anti-Spam Filtering", in *Proceedings of the Workshop on Machine Learning in the New Information Age, 11th European Conference on Machine Learning*, Barcelona, Spain, pages 9–17, 2000.

[3] Zhan Chuan *et al*., "An Improved Bayesian with Application to Anti-Spam Email", in *Journal of Electronic Science and Technology of China*, Vol.3 No.1, Mar. 2005.

[4] Vikas P. Deshpande and Robert F. Erbacher, "An Evaluation of Naïve Bayesian Anti-Spam Filtering Techniques", in *Proceedings of the 2007 IEEE Workshop on Information Assurance United States Military Academy*, West Point, NY 20-22 June 2007.

[5] Sheng, S. *et al*.,"An empirical analysis of phishing blacklists", in *Proceedings of the CEAS'09*, 2009.

[6] Pawan Prakash *et al*.,"PhishNet:Predictive Blacklisting to Detect Phishing Attacks", in *Proceedings of the IEEE Infocom*, pp.1-5, 2010.

[7] Maher Aburrous *et al*., "Experimental Case Studies for Investigating E-Banking Phishing Techniques and Attack Strategies", *Cognitive Computing*, DOI 10.1007/s12559-010-9042-7, Vol. 2, pp. 242-253, 2010.

[8] Congfu Xu *et al*.,"An approach to image spam filtering based on base64 encoding and N-Gram feature extraction", in *IEEE International Conference on Tools with Artificial Intelligence*, DOI 10.1109/ICTAI.2010.31, 2010.

[9] R. Malathi, "Email Spam Filter using Supervised Learning with Bayesian Neural Network", Computer Science, H.H. The Rajah's College, Pudukkottai-622 001,Tamil Nadu, India, Int J Engg Techsci Vol 2(1),89-100, 2011.

[10] Sadeghian, A and Ariaeinejad, R., "Spam detection system: A new approach based on interval type-2 fuzzy sets", in *IEEE CCECE -000379*, 2011.

[11] Xiang, G. *et al*., "CANTINA+: A feature-rich machine learning framework for detecting phishing Web sites". in *ACM Trans. Inf. Syst. Secur*. Vol.14, No.2, pp.1-21, 2011.

[12] Naïve Bayes Classifier.(2014, Dec) [online] Available : http://www.statsoft.com/textbook/naive-bayes-classifier .