SHED: Spam Ham Email Dataset

Upasana Sharma
Student
Central University of Punjab
Bathinda, Punjab
upasana.upasana06@gmail.com

Surinder Singh Khurana
Assistant Professor
Central University of Punjab
Bathinda, Punjab
surinder.seeeker@gmail.com

Abstract—Automatic filtering of spam emails becomes essential feature for a good email service provider. To gain direct or indirect benefits organizations/individuals are sending a lot of spam emails. Such kind emails activities are not only distracting the user but also consume lot of resources including processing power, memory and network bandwidth. The security issues are also associated with these unwanted emails as these emails may contain malicious content and/or links. Content based spam filtering is one of the effective approaches used for filtering. However, its efficiency depends upon the training set. The most of the existing datasets were collected and prepared a long back and the spammers have been changing the content to evade the filters trained based on these datasets. In this paper, we introduce Spam Ham email dataset (SHED): a dataset consisting spam and ham email. We evaluated the performance of filtering techniques trained by previous datasets and filtering techniques trained by SHED. It was observed that the filtering techniques trained by SHED outperformed the technique trained by other dataset. Furthermore, we also classified the spam email into various categories.

Keywords-*Spam Email, Non-spam emails, WEKA, feature selection, classifiers, Parameters*

*****

I. INTRODUCTION

Spam Email also known as junk email or unsolicited bulk email (UBE) is a subset of electronic spam involving nearly identical messages sent to numerous recipients by email[1]. The messages may contain disguised links that appear to be for familiar websites but in fact lead to phishing web sites or sites that are hosting malware[2]. Spam email may also include malware as scripts or other executable file attachments. Spammers collect email addresses from chat rooms, websites, customer lists, newsgroups, and viruses which harvest users' address books, and are sold to other spammers. **Ham email** is a term opposed to spam messages. Ham is then all "good" legitimate email messages, that is to say, all messages solicited by the recipient through an opt-in process[3].

II. LITERATURE REVIEW

There are large amounts of researches had been done by researchers to defend against spam. To enable the researchers various email datasets were also prepared. **SpamAssassin** is a public mail corpus[4]. This is a selection of mail messages, suitable for use in testing spam filtering systems. **TREC's** Spam [5] track uses a standard testing framework that presents a set of chronologically ordered email messages a spam filter for classification. In the filtering task, the messages are presented one at time to the filter, which yields a binary judgment (spam or ham). Four different forms of user feedback are modeled: with immediate feedback, delayed feedback, partial feedback and full delayed feedback. **Ling-Spam**[6] is a mixture of spam messages, and legitimate messages sent via the Linguist list, a moderated and spam-free mailing list about the science and profession of linguistics. The corpus consists of 2893 messages:2412 legitimate messages, obtained by randomly downloading digests from the list's archives, 481 spam messages, received by Ion Androutsopoulos, one of the authors of the corpus. Ling-Spam has the disadvantage that its legitimate messages are more topic-specific than the legitimate messages most users receive. Hence, the performance of a learning-based anti-spam filter on Ling-Spam may be an over-optimistic estimate of the performance that can be achieved on the incoming messages of a real user, where topic-specific terminology may be less dominant among legitimate messages. **Enron corpus**[7] is another email dataset. Authors analyzed its suitability with respect to email folder prediction, and provide the baseline results of a state-of-the-art classifier (Support Vector Machines) under various conditions, including the cases of using individual sections (From, To, Subject and body) alone as the input to the classifier, and using all the sections in combination with regression weights. A large set of email messages, the Enron corpus, was made public during the legal investigation concerning the Enron corporation.

Table 1: List of Spam Email Datasets

Corpus Name	Number of Messages		Year of Creation	Reference/Used
	Spam	Ham		
SpamAssassin	**1897**	**4150**	**2002**	[M´endez et al, 2006]
Enron-Spam	**13496**	**16545**	**2006**	[Koprinska et al, 2007]
LingSpam	**481**	**2412**	**2000**	[Sakkis et al, 1512]
PU1	**481**	**618**	**2000**	[Attar et al, 2011]
Spambase	**1813**	**2788**	**1999**	[Sakkis et al, 1512]
Trec	**52790**	**39399**	**2005**	[Androutsopoulos et al, 2000a]

III. RESEARCH METHODOLOGY

Methodology used to carry out the work is divided into five phases:

1) *Phase-1:Creation of various email address*

In phase 1, 80 email addresses are created to collect various spam as well as non spam emails in it. All the emails are created on gmail.com. Fictitious email addresses have been assigned to these email ids.

2) *Phase-2: Circulation of Email Address*

In the next step we populate these email addresses on the Internet to make them publically available. The aim of this activity is derived from the methodology followed by spammers to collect and prepare email list. We mentioned these email ids on various blog and registered these email ids on various sites such as sites belong to Online Gambling. We have a strong opinion that such kind of sites are not only providing activities like gambling but also using the user data like email addresses for other purposes. These purposes may include advertising and/or sending malicious or social engineering emails. Email addresses and websites on which these emails are registered/circulated are mentioned in table 2.

Table 2: Various websites where email addresses are registered.

1	Online Lottery
2	online rummy websites
3	online gambling
4	online jobs
5	freelancer work
6	online viagara
7	weight loss
8	online ipad
9	loan websites
10	Jeevansathi
11	online pizza
12	online viruses

3) *Phase-3: Feature Extraction*

In this phase various features of the emails has been extracted and stored in excel sheet. These features are Subject, body, recipient email id, sender email id, date, time and message id.. Features of 6002 emails have been obtained in this activity.

4) *Phase-4: Calculation of attributes*

Further we calculated the values of attributes used by existing spam detection techniques.

1. Subject Length
2. Body Length
3. Frequently Appearing Word List for Spam as well as Ham emails
4. Word Frequency for Spam as well as Ham emails.

B. *Phase 5: Classification*

Further using above dataset we evaluated the performance of the dataset by using various classifiers (Adaboost, Naïve bayes, Bayesnet, RandomForest)[8] to classify email dataset into spam and ham categories. We divided the corpora into proportions of ham (legitimate) and spam messages[9]. The experiment is performed with the 135 most frequent words in spam email and 154 most frequent words in ham email. To compare the efficiency of the dataset we also classified the dataset with list of 55 words specified in Spambase dataset and with list having most frequent 55 words prepares from the dataset.

Table 3: List of top 135 spam words

Get	most	Even	Much	Subject
Out	order	Work	marketing	Next
One	Per	First	Cost	Income
More	e-mail	Take	Easy	Save
Only	Name	Way	home	Service
People	Internet	Credit	message	Month
Money	Call	Offer	Full	Sent
Business	addresses	Web	sales	Apply
Over	Receive	See	Few	special
Email	Know	Million	place	Amount
Make	program	Search	Pay	Earn
Time	Now	Company	products	Limited
Just	Need	Software	show	Click
Like	Very	Number	hours	Advertisement
Free	Please	Best	great	Detail
information	address	Product	good	Here
New	Site	Right	available	Price
Send	Own	Give	check	Important
Mail	Use	Help	Less	Loan
List	Find	Start	Buy	Tax
future	PM	ID	version	Science
Case	Enron	Trade	Excel	Users
Daily	Field	FEE	content	ever
High	Read	Mobile	Dear	follow
Article	Try	Weekly	Such	Daily
account	Health	Based	directly	Rights
Parts	Receiving	Other	Reply	Basis

Table 4: List of top 154 non- spam words

Market	Junior	University	Commission	Microsoft
sure	Manager	Corporation	Holder	Sep
Letter	Details	Staff	Trainee	Mar
Computer	Project	Naukri	General	National
2017	Unsubscribe	Mountain	Deputy	Life
Last	Application	Graduate	Degree	Reading
Date	Image	View	.Net	Office
2016	Research	United	Advt	Here
PST	Subscribed	Technician	Create	Technology
Assistant	India	Powered	Development	Core
Updated	Engineer	Options	ASP.NET	DATES
Votes	Medical	States	Public	Singh
AM	Google	Torronto	State	IMPORTANT
Various	Bank	Form	Horoscope	Ghansham
Invites	Senior	Build	Multiple	Facebook
Officer	Code	News	Forwarded	Police
Ago	Institute	Client	Wipeout	August
Yesterday	Technical	Data	Blogs	Open
Entry	Official	Details	Executive	Tips
Stop	Group	Pradesh	Ontario	Operator
wrote	Library	Continue	SQL	Gmail
Results	Android	Code	Video	Charting
Scale	Selection	Bathinda	Delhi	Sarkari
API	District	Board	Chief	Horoscope

 ISSN: 2321-8169
1078 – 1082

HTML	Centre	Stenographer	Preparation	Framework
Division	Affairs	Punjab	App	Clerk
Singles	mailbox	Ferrand	Laboratory	Professor
C#	Bengal	Canada	Inspector	Government
Drive	Court	Newsletter	Applicants	Note
Unattended	Management	Correspondence	Tasking	Swapout
Examination	Insurance	Angular	Apprentices	

Furthermore, various feature selection approaches: Greedy Stepwise, Ranker Search, Best First and Genetic Search have been applied to find the reduce words in the list so that the detection time can be reduced.

IV. RESULTS AND DISCISSION

As per the methodology we collected the spam and ham emails. Total 6002 emails were collected. These email includes 4490 spam and 1512 ham emails. By manual analysis these spam emails have been categorized into 14 categories. These categories and number of emails we covered in each category are mentioned in table 5.Important point of note here is about "Article publication spam" category. In this category 348 out of 4490 spam emails fall, which may be indicating about commercialization of our education system.

Table 5: Categorization of spam emails

Types of spam email	**No. of spam emails**
Social media spam	298
Lucky winner spam	245
Dream job spam	401
Weight loss spam	335
Male enhancement spam	303
Debt Reduction spam	239
Credential spam	327
Loan spam	331
Tax spam	269
Discount offer spam	451
Article publication spam	348
Software related spam	297
Product sale spam	340
money stealing spam	306
Total	**4490**

Table 6 Performance in 10-fold cross validation with new 55 attributes

Classifiers	**Accuracy**	**Time Taken to build model (sec)**	**Precision**	**Recall**
Naïve Bayes	73.87	0.17	0.89	0.33
Bayesnet	91.78	0.42	0.87	0.9
RandomForest	99.18	12.95	0.99	0.97
Adaboost	87.78	1.72	0.86	0.79

Performance analysis of 6002 emails (spam as well as non-spam)with 55 new attributes with chisquare evaluator and ranker attribute selection at 10 fold cross validation. From figure 1, correctly classified instances or accuracy obtained for Random Forest is 99.18% and time taken to build model is 12.95 seconds.

Table 6 shows performance analysis of 6002 emails (spam as well as non-spam) with 55 new attributes with **chisquare evaluator and ranker attribute selection at 10 fold cross validation.**

Table 6, shows the values for accuracy, time, precision and recall rate that are used for the comparison of four algorithms. It evaluates that random forest gives the best result among all the four algorithms. It gives highest accuracy and runs efficiently on large databases, whereas Bayes Net gives the second highest accuracy with least time to build the models.

Table 7 Performance in 10-fold cross validation with old 55 attributes

Classifiers	**Accuracy**	**Time Taken to build model (sec)**	**Precision**	**Recall**
Naïve Bayes	79.04	0.09	0.95	0.68
Bayesnet	89.91	0.29	0.89	0.94
RandomForest	94.84	10.17	0.94	0.96
Adaboost	88.35	0.84	0.91	0.89

Table 7 shows performance analysis of 6002 emails (spam as well as non-spam) with 55 old attributes with **chisquare evaluator and ranker attribute selection at 10 fold cross validation.**

 ISSN: 2321-8169
1078 – 1082

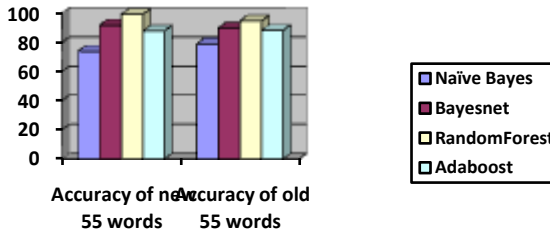

Fig1 Accuracy of different classifiers

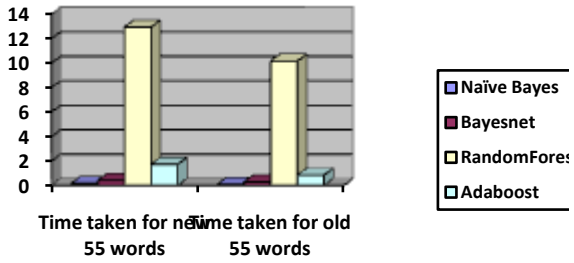

fig 2 Time taken to build models of different classifiers

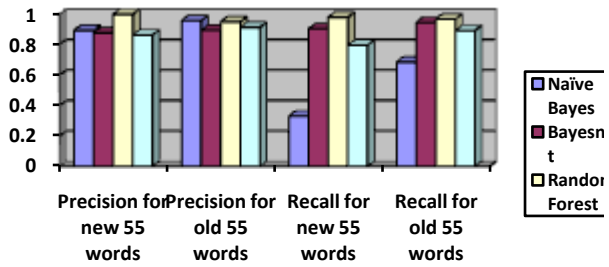

Fig 3 Precision and recall rate of different classifiers

Fig 4.4 shows performance analysis of 6002 emails (spam as well as non-spam) with 55 new attributes and old attributes with **chisquare evaluator and ranker attribute selection at percentage split (66:34).).** From fig 4, correctly classified instances or accuracy obtained is 98.68% and in fig 5 time taken to build model is 12.43 seconds.

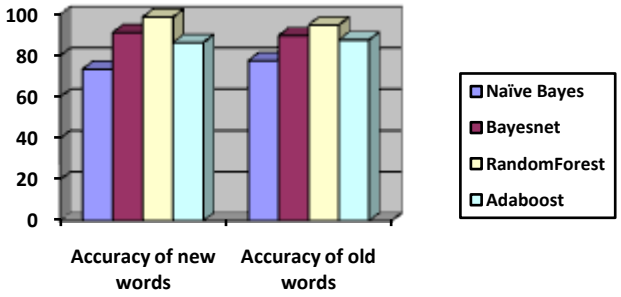

Fig 4 Accuracy of different classifiers

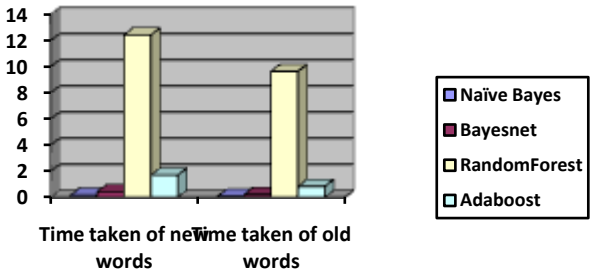

fig 5 Time taken to build models of different classifiers

V. CONCLUSION AND FUTURE WORK

In this work we prepared a dataset called SHED: Spam Ham Email Dataset. For the purpose various emails ids were created and circulated to public web domain so that the same can be collected by spammers. Around 6000 emails were collected to prepare the data set. Further we extracted various features from the emails. These features include: subject, body, sender's address, and receiver's address, email header, etc. Further we also calculated word frequencies. We prepared a list of frequently occurring words in both spam and non spam emails. To evaluate the spam email filtering process we use four classification approaches: Naïve Bayes, Bayesnet, Adaboost and RandomForest. These approaches were trained with SHED and spambase dataset(already existing). During the evaluation it was observed that the accuracy of all the classifiers is comparatively high while trained with SHED, as compared to the accuracy while trained with spambase. This is due to the change of different words used by spammers. The efficient among these classifiers is Random forest which classified the emails with 99.18% accuracy.

From the study, it can be concluded that spam cannot be completely stopped but it can be filtered with high accuracy. In future, several other algorithms other than the incorporated ones can be used to filter emails. Furthermore, work can be carried out to filter image based spam emails.

VI. REFERENCES

[1] P. Ozarkar and M. Patwardhan, "Efficient Spam Classification by Appropriate Feature Selection," *Glob. J. Comput. Sci. Technol. Softw. Data Eng.*, vol. 13, no. 5, 2013.

[2] S. Martin, B. Nelson, and A. D. Joseph, "Analyzing Behaviorial Features for Email Classification," *Micro*, vol. 3, pp. 123–133, 2004.

[3] S. Teli and S. Biradar, "Effective Spam Detection Method for Email," *Int. Conf. Adv. Eng. Technol.*, vol. 2014, pp. 68–72, 2014.

[4] A. A. Blaheta, "Project 3 : Spamassassin," no. November, pp. 2–7, 2014.

[5] G. Cormack and T. Lynam, "TREC 2007 Spam Track Overview," *Sixt. Text Retr. Conf. (TREC 2007)*, no. Trec, pp. 1–9, 2007.

[6] G. Costa, M. Errecalde, U. Nacional, D. S. Luis, and S. Luis, "Learning to detect spam messages."

[7] B. Klimt and Y. Yang, "The Enron Corpus: A New Dataset

for Email Classification Research," pp. 217–226, 2004.

[8] S. Sharma and A. Arora, "Adaptive Approach for Spam Detection," *Int. J. Comput. Sci. Issues*, vol. 10, no. 4, pp. 23–26, 2013.

[9] R. Sharma and G. Kaur, "Spam Detection Techniques : A Review," vol. 4, no. 5, pp. 2352–2355, 2015.

[10] Rathi and Megha and Pareek, Vikas,"Spam mail detection through data mining-A comparative performance analysis," Int. J. Comput. Sci. ,2013

[11] Trivedi, shrawan kumar and Dey,Shubhamoy,"A combining classifiers approach for detecting email spams" Proceedings IEEE,2004.