A MULTI-PRONGED APPROACH TO PHISHING EMAIL DETECTION

A Thesis Presented to

the Faculty of the Department of Computer Science

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

By

Nirmala Rai

December 2015

A MULTI-PRONGED APPROACH TO PHISHING EMAIL DETECTION

Nirmala Rai

APPROVED:

Dr. Rakesh M Verma, Committee Chairperson
Dept. of Computer Science

Dr. Arjun Mukherjee
Dept. of Computer Science

Dr. Chris R Bronk
Dept. of Information and Logistics Technology

Dean, College of Natural Sciences and Mathematics

Acknowledgements

It has been a fulfilling journey as a Master Thesis student and during this journey, I learned many things, interacted with many individuals, worked on many solutions and improvised for many problems. I would like to take a moment and thank each person who graciously extended help to me and contributed towards my success.

My academic advisor Dr. Rakesh Verma has truly been a friend, philosopher and guide throughout my academic career and I express my whole-hearted gratitude towards him for all his support and encouragement. He has ignited in me a passion for research, that will keep me curious and focused for years to come. I would like to thank Dr. Arjun Mukherjee and Dr. Chris Bronk for accepting the role and responsibilities as my thesis committee members. Their valuable suggestions and guidance helped in shaping my thesis well and improving it further.

Being a part of a very dynamic research lab ReDAS (Reasoning and Data Analysis for Security), was a privilege and I thank all my fellow ReDASers for their suggestions, ideas and co-operation. I would specially like to mention Arthur Dunbar, Avisha Das, Keith Dyer, Luis Felipe and Vasanthi Vuppuluri for their support.

Family and friends can be the strongest anchor in our life and have a great contribution in our achievements. I sincerely thank my parents B.L Rai and Kamala Rai for believing in me, my brother Narendra Rai for having confidence in me, even more than myself and my close friend Chirag Chatterjee for being with me and motivating me through the toughest of times. Lastly, I offer my deepest humility to the Almighty for all that I have accomplished.

This research is supported in part by NSF grants CNS 1319212 and DUE 1241772.

A MULTI-PRONGED APPROACH TO PHISHING EMAIL DETECTION

An Abstract of a Thesis

Presented to

the Faculty of the Department of Computer Science

University of Houston

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

By

Nirmala Rai

December 2015

Abstract

Phishing emails are a nuisance and a growing threat for the world causing loss of time, effort and money. In this era of online communication and electronic data exchange, every individual connected to the Internet has to face the danger of phishing attacks. Typically, benign-looking emails are used as the attack vectors, which trick users into revealing sensitive information like login credentials, credit-card details, etc. Since every email contains important information in its header, this thesis describes ways of capturing this information for successful classification of phishing emails. Moreover, the phisher has total control over the email body and subject, but little control over the header after the email leaves the sender's domain, unless the phisher is sophisticated and spends a lot of time crafting the attack, which reduces the payoff or may even backfire or yield mixed results.

This thesis is a consolidated account of various systems designed to combat phishing emails from different dimensions. The main area of focus is email header. Techniques like n-gram analysis, machine learning and network port scanning are used to extract useful features from the emails. This thesis shows that the classes of features used in these systems are very effective in distinguishing the phishing emails from the legitimate ones. Using different real datasets from varied domains, it highlights the robustness of the methods presented. Some methods, like the header-domain analysis, obtain high detection rates of 99.9% and low false positive rates of 0.1%. These approaches have the advantage and flexibility that they can be easily combined with other existing methods, in addition to being used in standalone mode.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

As the Internet has become an integral part of our lives we have entered an age of online transactions. Almost every service we use has an online access portal. These services need payment, for which we provide sensitive information like credit-card details, bank account numbers etc. Since these websites handle such sensitive information, their login credentials are also equally private and sensitive. Phishing refers to the act of attempting to steal valuable and sensitive data from individuals and organizations. It causes huge monetary and information loss. Such phishing attacks are targeted towards obtaining valuable information from users as mentioned above.

1.1 Motivation

The APWG Phishing Activity Trends Report for the 4th quarter of 2014 [4] showed that the number of unique phishing reports submitted was 197,252, 18% more than in the 3rd quarter. The number of unique phishing sites had also increased from 14,258 in November, 2014 to 17,320 in December, 2014. A total of 437 brands were targeted in Q4 and United States was the country hosting the greatest number of the phishing sites, yet again. This clearly shows that phishing detection is still an unsolved problem and one which causes heavy damage to the people and the society.

The most common means used by the phishers are benign-looking emails that lure users and trick them into revealing the sensitive data and thus result in loss and misuse of valuable private information apart from monetary losses. Since email is the most commonly used channel for phishing attacks, this thesis concentrates on working towards an effective way of segregating them into phishing and legitimate classes.

Every email consists of two parts: the header and the body. The header consists of several pre-formatted fields such as From, Delivered-To, Subject, Message-ID, etc. The body consists of the main content of the email, usually in text/HTML format. The phishers make it very difficult to detect the phishing emails by meticulously constructing them to closely resemble legitimate ones. This makes the process of distinction non-trivial, which has been observed by other researchers [23] also.

The email body is completely under the control of the sender while the header follows a relatively stricter format and is not entirely controlled by the sender. So the

main focus of this thesis is on detection based on email headers, with some attention to the email body text as well. In particular, based on looking at a few (less than 10) legitimate emails and the same number of phishing emails, our attention was drawn to the Message-ID in the header fields. This field is a string following a certain basic format described in the Background Chapter. It also contains information designed to make the email globally unique. It cannot be altered easily and it provides important information about the email that includes it.

A part of the work presented here centers on these useful properties of Message-IDs and exploits it further by applying n-gram analysis to the Message-IDs. Various machine-learning algorithms including an on-line confidence weighted-learning algorithm were employed using 10-fold cross validation on different data sets and they produced detection rates of above 99%. To our knowledge, this is the first time Message-IDs have been used with n-gram analysis to detect phishing emails. This system is named Phish-IDetector and is explained in details in Chapter 5.

The headers were also observed to contain several domain names. These domains contained information which can be used to trace the path of the email. The domains closer to the receiver's side cannot be altered easily and it provides important information about the trail the email has followed. Different systems were built to extract features from these domains for classification which have been described in Chapters 6, 7 and 9.

1.2 Overview of Thesis

A brief overview of the rest of the chapters in this thesis is mentioned below.

1. Chapter 2: Includes background knowledge and preliminary information required to get a better understanding of the thesis.
2. Chapter 3: Includes description of Phish-IDetector, a Message-ID based phishing detection system.
3. Chapter 4: Includes details about a combined grand experiment using header and text analysis.
4. Chapter 5: Includes the description of email header domain analysis.
5. Chapter 6: Includes the description of SMTP analysis for emails.
6. Chapter 7: Includes study of the domain details obtained from the different datasets.
7. Chapter 8: Includes the description of path analysis for emails.
8. Chapter 9: Includes the relevant related work for this thesis.
9. Chapter 10: Concludes the thesis.

1.3 Contribution

The major contributions of this thesis are as follows:

1. The demonstration that Message-ID and domains from email header fields are effective in phishing email detection [Chapter 3 and 5].

2. The approach of applying n-gram analysis technique with a rich variety of classifiers to these email header properties [Chapter 3 and 5].

3. A novel approach to path analysis of emails by reconstructing the route using Received-From and 'by' pairs [Chapter 8].

4. The SMTP experiment whereby we checked for open SMTP servers as an indication of use of source routing by the phishers [Chapter 6].

5. A preliminary study of the domain details obtained from the different datasets [Chapter 7].

Chapter 2

Background

Electronic mail or email proliferated during the 1990s and has evolved to become an indispensable part of our current social fabric. Essentially, an email has two parts: the header and the body. The email body contains the actual message being sent and is completely under the control of the sender. Whereas, email header consists of several fields, some mandatory and some optional, which carry information regarding the source, destination, routing details, timestamps, etc. [37]. Thus, the header cannot be completely manipulated by the sender.

Every email contains information of the path it has taken since it left the sender's mail box till it reaches the receiver's mail box. This information can be extracted from the header fields of the email. Let's consider a simple example where sender A sends an email to receiver B. Though there are many header fields we will be focusing on the ones which concern the email's path of travel. In the most basic form, there are four entities involved. The sender's mail client referred to as a@sender.com, sender's

mail server mail.sender.com, receiver's mail server mail.receiver.com and receiver's mail client b@receiver.com. The email is created by the sender using his mail client and contains his own address as 'From' header and the receiver's email as the 'To' header. Once sent, it passes on to his mail server which adds some header fields like the 'Received: from' and the 'Message-ID' headers. Similarly, the receiver's mail server adds 'Received: from' header before passing it on to the receiver's mail client. Hence, at the end of the delivery process the length of email's header increases with every hop. In a more complex scenario where the email passes through several Mail Transfer Agents (MTAs), more header fields are added to it.

We provide an explanation of some terms we will use frequently throughout the paper.

Header domains. The header fields concerned with the transfer and delivery of the emails mostly contain the name of the domain of each mail client and server that it passes through. For example in the 'From' field address a@sender.com, 'sender.com' is the domain name. These domains are extracted from all the fields which contain such information and are collectively addressed as 'Header domains' in the paper.

Message-ID. RFC 2822 [37] is a standard that specifies the syntax for messages that are sent as "electronic mail" messages. It states that each email should have a globally unique identifier called Message-ID. If this is included, it must be in the email header. RFC 2822 also defines the syntax of Message-ID. It should be like a legitimate email address and it must be included within a pair of angle brackets. A typical Message-ID looks like the following:<20020923025816.8E7A34A8@mercea.net>. According to RFC 2822, Message-ID can appear in three header fields. They are (i)

Message-ID header, (ii) In-Reply-To header and (iii) References header. The "In-Reply-To:" and "References:" fields are used while creating a reply to a message. They hold the message identifier of the original message and the message identifiers of other messages (e.g. replies to the message). The "In-Reply-To:" field may be used to identify the message (or messages) to which the new message is a reply, while the "References:" field may be used to identify a "thread" of conversation [37]. But Message-ID of the present email should be included against the Message-ID header [32]. The Message-ID has a fixed format of the form <LHS@RHS>where the left hand side (LHS) is a representation of information including current time stamp, queue id, etc. coded in different formats according to the Sendmail version. The right hand side (RHS) represents the *fully qualified domain name* (FQDN). This part starts with local host name followed by a dot and other parts of domain information [12].

N-gram. The concept of N-gram is related to natural language processing. It is a sequence of n characters in a string or text. For example if the text is abc123 the 1-grams would consist of one character sequence, e.g., a, c, 2, etc. Similarly, 2-grams would be overlapping sequence of 2 characters like ab, bc, c1, 12, 23. This idea can be further extended to higher order n-grams in a similar fashion.

Chapter 3

Phish-IDetector

[This chapter's contents have been published in 12th International Conference on Security and Cryptography [42], SECRYPT 2015.]

Sendmail, one of the Mail Transfer Agents (MTAs) uses Message-ID for tracing emails and for logging process ids [12]. Sendmail specification recommends including Message-ID in emails and also the setting of relevant macros in its configuration file in order to implement compulsory checking of Message-IDs [12]. "Unlike spoofing other fields in the header, spoofing Message-ID needs special knowledge. Only technical savvy spammers can spoof the Message-ID cleverly" [32]. So, deep analysis on Message-IDs may reveal some sort of information that could open a window to trace the source of an email.

3.1 Overview

Based on the above hypothesis, we delved deeper into Message-ID using n-gram analysis up to 10-grams and found the optimum detection rates at around 5- or 6-grams. For both higher and lower order n-grams the rates usually deteriorate. We applied several machine learning classifiers using stable version 3.6 of Weka [19] and an on-line confidence weighted learning algorithm of [28]. The complete process can be summarized in the following sequence of steps.

3.1.1 Message-ID Extraction

For our study we decided to choose Message-ID as the distinguishing property because of its content, uniqueness and fixed format. All the Message-IDs from the emails of different datasets are extracted using grep command and stored in a file. Since each Message-ID is of the format <LHS@RHS>, we get rid of the <, @ and > symbols common to all Message-IDs as a pre-processing step. After this step we get two attributes for each Message-ID. We have named them LHS and RHS to denote the left hand side and the right hand side of the Message-ID.

3.1.2 Input File Creation

Depending on which dataset the email belonged to, we labeled each instance as belonging to either "phishing" or "legit" (legitimate) class. We created a csv file with three columns: class label, LHS and RHS. The RHS part being of a more

consistent format rather than LHS, we tried the classification based on only RHS as well. In that case, there are only two columns: class label and RHS.

3.1.3 N-gram Analysis

Further, we performed n-gram analysis of the collected Message-IDs so that we could represent the data in numeric format acceptable to most classifiers in Weka. We decided to use n-gram analysis, as this kind of analysis is able to capture the structure present in any text or string. As discussed in [9] the main advantage of N-gram-based analysis is in its nature of n-gram creation. It helps in minimizing errors and limiting it to only the n-grams derived from the erroneous part because every string is decomposed into small parts. The remaining part of the text remains error-free. The count of common n-grams between two strings is a good measure of text similarity, and this measure has proved to be resistant to different kinds of textual errors. This analysis generates unique n-gram features represented as Unicode code point of the characters. For example, the Unicode code point for "a" is 97 so the 1-gram "a" will be represented as 97. The feature extracted is the frequency of the n-gram in the attribute LHS or RHS.

3.1.4 Classification

Once we obtained and represented the data in arff format, we ran the following six classifiers on the arff file using Weka 3.6: RandomForest [8], J48 [36], Bagging [7], AdaboostM1 [17], SMO [33] and NaiveBayes [24]. We also ran four other classifiers

but their performance was not at par with the 6 classifiers mentioned, so we omit their results. These are: ClassificationViaClustering, ComplementNaiveBayes, ZeroR and BayesNet. The arff files were also converted to .svm format, the input format for the confidence weighted algorithm [28], using a python script.

3.2 Data Sets and Classifiers

We have used two publicly available datasets. Email Message-IDs were collected from 4,550 public phishing emails from [30] and from 9,706 legitimate emails from SpamAssassin public corpus datasets at [3]. The phishing corpus and even the SpamAssassin ham corpus we used has been used previously by [16], [38], [20].

SpamAssassin corpus segregates the emails into different subsets which we named as follows:

1. easy_ham consisting of 5051 emails
2. easy_ham_1 consisting of 2500 emails
3. easy_ham_2 consisting of 1400 emails
4. hard_ham consisting of 500 emails
5. hard_ham_1 consisting of 250 emails

All these emails had Message-IDs. We ran experiments on the phishing emails combined with each of the above mentioned subsets of legitimate emails. All the

Message-IDs obtained from the phishing emails were added to the set of Message-IDs extracted from each of the above mentioned ham sets. These datasets are hence named according to the ham set involved in creating the dataset since the phishing set of Message-IDs is common to all of them. Additionally, all the experiments were performed once taking only RHS into account and once taking both the RHS and LHS into account, and the dataset names have been prefixed with RHS and SplitMsgId respectively. The names of the datasets are as follows:

1. RHSEasyHam and SplitMsgIdEasyHam
2. RHSEasyHam1 and SplitMsgIdEasyHam1
3. RHSEasyHam2 and SplitMsgIdEasyHam2
4. RHSHardHam and SplitMsgIdHardHam
5. RHSHardHam1 and SplitMsgIdHardHam1

We used Weka version 3.6 which is basically a collection of machine learning algorithms for data mining tasks. It was chosen because of its wide acceptability, popularity and its ease of use. It has previously been used for phishing detection by [20] and [11]. Weka provided us an easy method of comparing the performance of several classifiers on our datasets and choosing the best among them. We ran the experiments with around 10 classifiers and chose the best 6 among them. Each of them is explained here in brief.

Random Forest classifier (RF) [8] consists of several decision tree classifiers. Each tree has a random set of features out of the total feature collection and this

algorithm returns the maximum frequency class among all of the individual decision trees. It performed the best quite consistently in our experiments. For our experiments we used the default implementation of Weka 3.6 for the Random Forest classifier.

J48 is a Java implementation of the decision tree formed by classifier C4.5 [36].

SMO is an implementation of sequential minimal optimization algorithm devised by John Platt for training a support vector classifier. All attributes are normalized by default in this algorithm. A more detailed explanation can be found at [33].

Bootstrap Aggregating or **Bagging** is a method for generating multiple versions of a predictor and using these to get an aggregated predictor. The aggregation averages over the versions when predicting a numerical outcome and does a plurality vote when predicting a class. It is explained in [7].

AdaBoostM1 (ABoost) is an implementation of the boosting algorithm by [17]. It is known to improve performance of a weak learner using a boosting algorithm. We have used the default base classifier for Weka 3.6 in this case.

NaiveBayes (NB) is the Weka implementation of the Naive Bayes classifier, which is a simple classifier that applies Bayes' theorem. It strictly assumes conditional independence and hence called 'naive.' More information can be found at [24].

For classification based on higher order n-gram analysis we used the faster online confidence weighted learning algorithm of [28]. We obtained a collection of most confidence weighted learning algorithm into a library written in Java from [13].

Again, we selected the 10-fold cross-validation test option for maintaining uniformity. With this algorithm we were able to perform the classification for all the files up to 10-grams.

3.3 Independent Experiment on Message-IDs

Due to privacy issues, legitimate emails used in the field of phishing emails detection are usually not recent ones. To prove the viability of our method with current data without compromising the privacy aspect, we performed an independent experiment involving 10 anonymous volunteers. Each of them was given instructions along with a script that would collect some statistics from each mail box. We collected only two numbers from each of them, no. of emails (Email Count) and number of Message-IDs (Message-ID Count) not revealing any private data in their emails. The process involved configuring each volunteer's gmail account in their local UNIX machines using postfix and fetchmail. The script then separated the mailbox created for each volunteer into individual messages using procmail. And finally grep command was used to get the email count and the Message-ID count. We had to be careful not to over count the Message-IDs as sometimes a mail can have more than one Message-ID. To avoid such a mistake we used grep with the option of counting only the first occurrence of Message-ID in each email.

The data collected from the volunteers is shown in Table 3.3. It reveals that nearly 99% of the emails have Message-ID field and proves our hypothesis that in spite of being an optional field it would have to be included in the emails by a phisher

to avoid raising any red flags.

Table 3.1: Email and Message-ID count from the independent experiment. Nearly 99% emails have Message-Ids.

Message-ID Experiment		
Volunteers	**EmailCount**	**Message-ID Count**
Volunteer 1	1959	1928
Volunteer 2	1613	1594
Volunteer 3	798	787
Volunteer 4	719	712
Volunteer 5	364	361
Volunteer 6	352	352
Volunteer 7	325	325
Volunteer 8	277	263
Volunteer 9	252	252
Volunteer 10	118	118
Total	6777	6692
Percentage Emails With Message-IDs		
98.75		

3.4 Results

Since the file-size increased exponentially for each subsequent n-gram, Weka would crash for any n-gram higher than 3. Also, we could run only two classifiers for the 3-grams files due to the issue of large-sized files. For both 1- and 2-grams files we ran as many as 10 classifiers and found Random Forest to be the most effective of them all, obtaining highest True Positive rate (TPR) and the lowest False Positive Rate (FPR).

TPR refers to the percentage of instances of a class x, classified correctly among all the instances truly having the class x, i.e., what part of the class was captured.

FPR refers to the percentage of instances of a class x classified incorrectly as some other class among all the instances not of class x [34].

Looking at both the TPR and FPR values of these experiments, it was revealed that with an increase in order of n-gram, the classification improves but it starts deteriorating after a certain n-gram value. For most of the experiments this optimum value was obtained at the threshold of around 5- or 6-grams.

We present our results of all classifiers for the best among all the datasets, i.e., Hard Ham. Also, to give an idea of the performance across all datasets we include the results of our best classifiers, i.e., Random Forest and J48 for all the datasets.

Tables 3.1 to 3.3 summarize the TPR and FPR values of the experiments on dataset SplitMsgIdHardHam. Results show a constant increase in TPR and decrease in FPR for higher order n-grams. So, the 3-grams results are the best in terms of both TPR and FPR. Random Forest classifier even succeeds in getting 99.5% of the phishing emails detected with a small number of false positives, i.e. legitimate emails classified as phishing.Also, we find that the classifiers that perform the best classification are tree classifiers Random Forest and J48.

Tables 3.4 to 3.6 summarize the TPR and FPR values of the experiments on dataset RHSHardHam. Similar to the SplitMsgIdHardHam dataset results there is a constant increase in TPR and decrease in FPR for higher order n-grams. So, the 3-grams results are the best in terms of both TPR and FPR.

Note that the SplitMsgIdHardHam dataset gives better results as compared to the RHSHardHam dataset. We hypothesize that many phishers lack adequate knowledge

of LHS structure or do not spend time on it. Both RandomForest and J48 perform almost consistently well for both the datasets at almost any n-gram value.

Tables 3.7 to 3.10 summarize the TPR and FPR values of the RandomForest and J48 classifiers for the experiments across all datasets. These two classifiers performed the best and we present these tables to compare their results for each of the datasets we used. We find that the results are fairly consistent across datasets and there is a gradual improvement of results with the increase in the order of n-grams.

We present the results of Confidence Weighted Classifier for all 10-gram datasets in figures 10.1 and 10.2. The advantage of Confidence Weighted algorithm was that it could easily run on all the 10-gram files and that it had quite low false positive rate consistently as compared to the Weka machine learning classifiers. Though the detection rates are not as high as Random Forest and J48 classifiers, the false-positive rates are much lower.

3.5 Information Gain

After the first set of experiments, we were curious to know which features were performing the best among all of the 1-gram, 2-gram and 3-gram attributes. A widely accepted method to find out the most effective features in a multi-feature classifier is calculating the information gain for the attributes. It is a measure of the difference in entropy values. We present the top 10 features along with their information gain values for each of the 1-gram, 2-gram and 3-gram features. From these IG values we find that for the RHSHardHam dataset, the hyphen '-' symbol is

quite dominant as an attribute.

3.6 Security Analysis

Our method relies on the Message-ID field which, though important and recommended, is optional. Without it, our method would not work. However, note that almost all legitimate emails include this field, and since a phisher tries to fool the user into believing that a phishing email is legitimate, omitting this field could serve as a red flag. In our experimental data set, 100% of the legitimate emails had Message-IDs. Our recent experiment with 10 volunteers reveals that the Message-ID field is present in nearly 99% of the legitimate emails. Also, the exponentially increasing file size for higher order n-grams makes it difficult to run different classifiers on them without using specialized big data approaches. We currently ran only the confidence weighted algorithm on higher order n-gram files, which has proven itself to be competitive in other scenarios, but not guaranteed to be the ideal choice for best results.

Spoofing Message-ID field requires a technically savvy phisher, who is willing to go the extra mile to avoid detection. For example, either this field would have to be deleted, which would raise a red flag in light of our experiment with 10 volunteers, or the phisher would: (i) either fake the FQDN or (ii) copy the entire Message-ID field from a legal message sent earlier, and the phisher would have to turn off any checking in the mail program. For such sophisticated phishers, we recommend combining our classifiers with other classifiers or features from the header, the links and the body

text in the email, as, for example, in [41].

Table 3.2: True-Positive and False-Positive Rates for Weka Classifiers on SplitMsgIdHardHam Dataset

1-gram for SplitMsgIdHardHam		
Classifiers	TPR	FPR
RandomForest	99.5	4.9
J48	96.6	18
Bagging	96.7	27.5
SMO	94.3	46.8
AdaboostM1	94.9	37.3
NaiveBayes	87.2	29.7

Table 3.3: True-Positive and False-Positive Rates for Weka Classifiers on SplitMsgIdHardHam Dataset

2-gram for SplitMsgIdHardHam		
Classifiers	TPR	FPR
RandomForest	99.4	4.9
J48	97	18.4
Bagging	97.2	23.2
SMO	97.6	8.8
AdaboostM1	95	37
NaiveBayes	92	29.1

Table 3.4: True-Positive and False-Positive Rates for Weka Classifiers on SplitMsgIdHardHam Dataset

3-gram for SplitMsgIdHardHam		
Classifiers	TPR	FPR
RandomForest	99.3	5.2
J48	98.7	8

Table 3.5: True-Positive and False-Positive Rates for Weka Classifiers on RHSHard-Ham Dataset

1-gram for RHSHardHam		
Classifiers	TPR	FPR
RandomForest	99.4	5
J48	96.5	17.6
Bagging	96.7	27.4
SMO	94.3	46.8
AdaboostM1	93	59.4
NaiveBayes	88.1	45.4

Table 3.6: True-Positive and False-Positive Rates for Weka Classifiers on RHSHard-Ham Dataset

2-gram for RHSHardHam		
Classifiers	TPR	FPR
RandomForest	99.3	5.2
J48	98	10.5
Bagging	97.7	18.6
SMO	98.8	5.5
AdaboostM1	93.9	54.1
NaiveBayes	92.4	35.8

Table 3.7: True-Positive Rate and False-Positive Rate for Weka Classifiers on RHSHardHam Dataset

3-gram for RHSHardHam		
Classifiers	TPR	FPR
RandomForest	99.4	5
J48	97.4	16.8

Table 3.8: True-Positive and False-Positive Rates for RandomForest and J48 across all SplitMsgId Datasets

1-gram				
DataSet	**RandomForests**		**J48**	
(Split)	TPR	FPR	TPR	FPR
EasyHam	93.7	10.1	90.2	12.7
EasyHam1	95.7	4.6	91.3	9.2
EasyHam2	95.4	13.2	91	16.4
HardHam	99.5	4.9	96.6	18
HardHam1	98.5	26.5	97.3	36.4

Table 3.9: True-Positive and False-Positive Rates for RandomForest and J48 across all SplitMsgId Datasets

2-gram				
DataSet	**RandomForests**		**J48**	
(Split)	TPR	FPR	TPR	FPR
EasyHam	93.9	9.9	91	12.1
EasyHam1	96.1	4.2	92.3	8.2
EasyHam2	95.9	12.7	92.9	15
HardHam	99.4	4.9	97	18.4
HardHam1	98.5	26.1	97.8	33

Table 3.10: True-Positive and False-Positive Rates for RandomForest and J48 across all RHS Data Sets

1-gram				
DataSet	**RandomForests**		**J48**	
(RHS)	TPR	FPR	TPR	FPR
EasyHam	95.6	4.7	91.4	9.1
EasyHam1	93.7	10.1	90.2	12.7
EasyHam2	95.3	13.3	91	16.4
HardHam	99.4	5	96.5	17.6
HardHam1	98.5	26.1	97.3	36.8

Table 3.11: True-Positive Rate and False-Positive Rate for RandomForest and J48 across all RHS Data Sets

2-gram				
DataSet	**RandomForests**		**J48**	
(RHS)	TPR	FPR	TPR	FPR
EasyHam	94.8	8.8	95.9	5.3
EasyHam1	98.5	1.5	96.6	3.4
EasyHam2	95	15.1	96.1	7.5
HardHam	99.3	5.2	98	10.5
HardHam1	97.9	36.4	98.4	22

Figure 3.1: True-Positive and False-Positive Rates for Confidence Weighted Classifier on SplitMsgIdHardHam Dataset

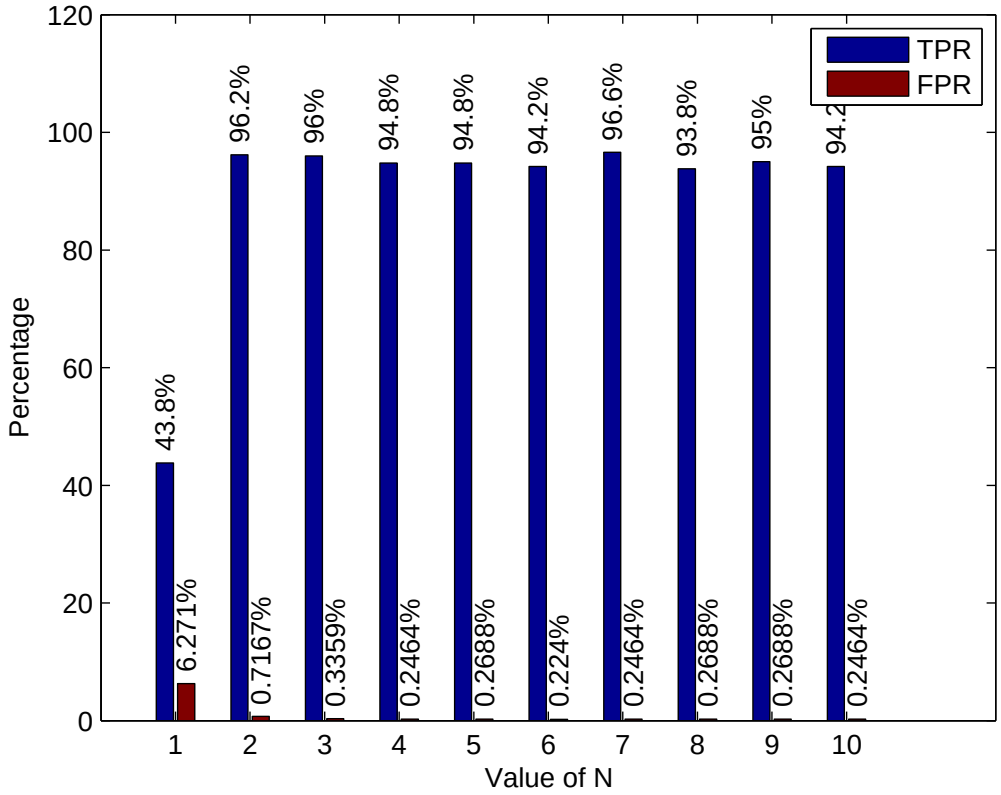


Figure 3.2: True-Positive Rate and False-Positive Rate for Confidence Weighted Classifier on RHSHardHam Dataset

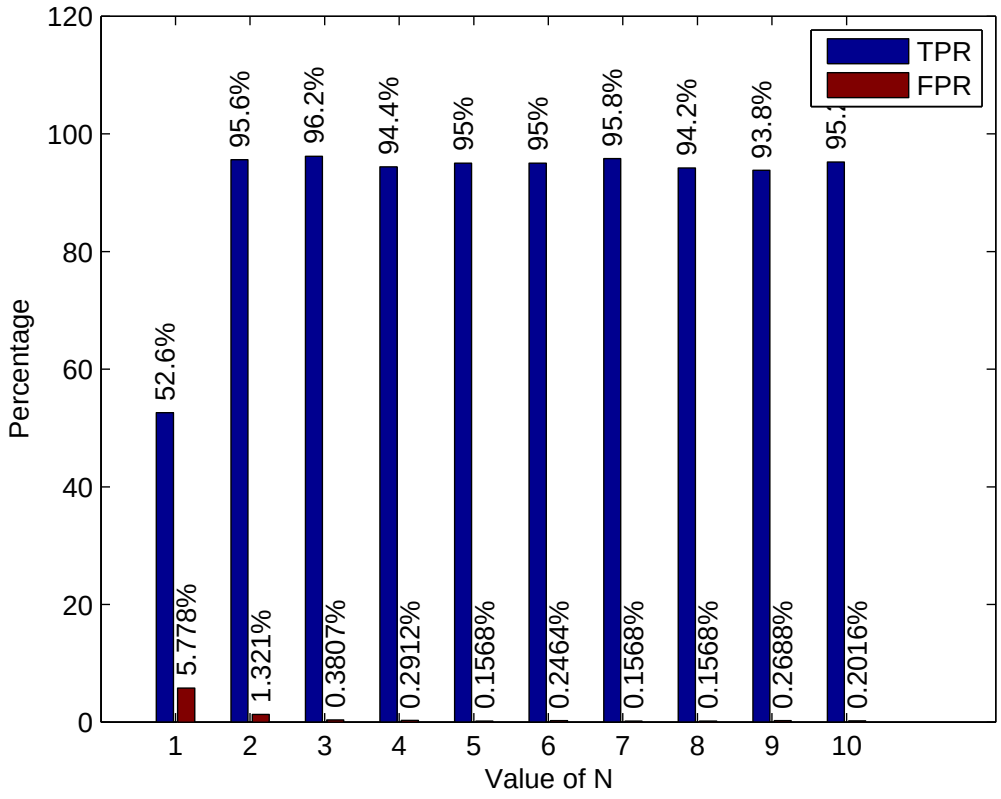


Table 3.12: Information gain values of Top 10 attributes represented as 'Att' in the table for RHSHardHam Data Set.

RHSHardHam					
1-gram		**2-gram**		**3-gram**	
Att	**IG**	**Att**	**IG**	**Att**	**IG**
-	0.110599	-a	0.125108	-sf	0.125992
a	0.103357	bv	0.118968	-a	0.125108
e	0.073969	sf	0.118461	-ac	0.122093
.	0.063256	v-	0.117757	fo1	0.122093
s	0.060871	1-	0.116663	-ag	0.122093
f	0.059647	o1	0.1156	abv	0.122093
t	0.058247	c-	0.11354	bv-	0.122093
g	0.055149	-	0.110599	o1-	0.122093
n	0.049335	-s	0.108526	1-a	0.122093
b	0.045831	a	0.103357	sfo	0.122093

Table 3.13: Information gain values of Top 10 attributes represented as 'Att' for SplitMsgIdHardHam Dataset

SplitMsgIdHardHam					
1-gram		**2-gram**		**3-gram**	
Att	**IG**	**Att**	**IG**	**Att**	**IG**
.	0.2007	.	0.200703	.	0.200703
a	0.16614	a	0.16614	a	0.16614
o	0.15855	o	0.158549	o	0.158549
t	0.10909	l.	0.146136	il.	0.146136
r	0.10193	.J	0.14211	l.	0.146136
l	0.09794	Ma	0.140517	.10	0.144005
i	0.09426	aM	0.139745	.J	0.14211
v	0.0927	av	0.139732	Ma	0.140517
M	0.06281	ot	0.139432	Mai	0.139745
-	0.05612	va	0.138928	vaM	0.139745

Chapter 4

Grand Experiment

4.1 Enhancement of Semantic Feature Selection

A general semantic feature selection method for text problems was proposed by [40] which is based on the use of statistical t-test and WordNet - a lexical database, that can work as both a dictionary and a thesaurus [15]. In the semantic feature selection method, the email body text was used for feature selection using t-test. Weight calculation of the features was done and the features with weights above a certain threshold were used to form appropriate sets. These sets of features were then used in different classifiers. Here is a brief overview of the system proposed.

The authors observed that 84.7% of the phishing emails had the word 'your', as opposed to 34.7% of the legitimate emails. So, all the bi-grams (sequence of 2

words) following the word ‘your’ were collected along with their frequencies and t-test was performed to choose appropriate bi-grams as features. Further weights were calculated for these chosen bi-grams and a final set was formed with the bi-grams having weights above the set threshold. This set was called PROPERTY. Similarly, they worked with all the words in sentences having a hyperlink or any word from the set: ‘url’, ‘link’, ‘website’. After t-test and weight calculation, the resulting set was called ACTION. The text in the subject field of the emails were also collected. The stopwords were removed from the subject and t-test was performed on the remaining words to select the features forming the set PH-SUB.

4.2 Preliminaries

Some terms used further in this chapter are closely related to Natural Language Processing (NLP). We describe them briefly here for a better understanding of the reader.

Word Sense refers to the particular sense or meaning of a word, among its different meanings, that is used in a particular sentence. It is important to understand the complete meaning of the sentence. For example, in case of the word ’bank’, it needs to be clear whether it means the financial institution or the sides of a river.

Named Entity refers to the parts of texts that are nouns belonging to different categories like person, place, organization, etc.

Hyponym simply means a more specific term. It is closely related to the concept of hypernym which refers to a more general term in terms of the meaning of the word. For example spoon is a type of or a more specific form of cutlery. Hence, spoon is the hyponym of cutlery and cutlery is the hypernym of spoon.

Their system consisted of four different classifiers:

Classifier 1: Pattern Matching (PM)

Classifier 2: PM + Part of speech (POS) Tagging

Classifier 3: PM + POS + Word Senses

Classifier 4: PM + POS + Word Senses + WordNet

Pattern Matching involved two subclassifiers: Action-detector and Nonsensical-detector. Action detector marked an email as phishing if it has: i) the word 'your' followed by a bi-gram belonging to PROPERTY (for example, 'your credit card'), and ii) a word from ACTION in a sentence containing a hyperlink or any word from set: 'url', 'link', 'website'. All of these words were selected irrespective of the cases, i.e. both upper and lower case versions were considered. Nonsensical detector checked if the email subject has at least a named-entity, or a word from PH-SUB. If so, the email was marked as phishing if i) it contains at least one link, and ii) its text is not similar to the subject.

They provided the definition of 'similar' as follows: An email body text is similar to its subject if all of the words in the subject (excluding stopwords) are present in the email's text.

PM + POS Tagging classifier build on the previous one by using part of speech tags. Bi-grams not containing a noun or a named-entity are removed before forming the set PROPERTY. Similarly, words that are not verbs are excluded from the analysis for set ACTION and only named-entities, nouns, verbs, adverbs and adjectives were used for making the set PH-SUB.

PM + POS Tagging + Word Senses extended the classifier 2 by including the senses of words using SenseLearner [29]. The statistical analysis was performed on words with their POS tags and senses.

PM + POS Tagging + Word Senses + WordNet is the final classifier and it works by extending the sets ACTION, PROPERTY and PH-SUB by computing the synonyms and the direct hyponyms of the synonyms of each selected feature in the sets. Wordnet is used to get the synonyms and the direct hyponyms in this case.

4.3 Enhancement

Apart from re-implementing the above mentioned system, some observations and minor enhancements were also made. First of all the selected bi-grams were the ones occurring after you, yours, your's so that the phishing emails having these variations of the word 'your' may also get detected. Also for the Action-detector, the sentences having the words: site and hyperlink were also included in the analysis process.

We wanted to study the combined effect of different classifiers so we performed a grand experiment involving the following classifiers:

4.4 Semantic Feature Selection Pattern Matching

Pattern Matching as explained above.

4.5 PhishNet-NLP Enhanced Header Analysis

The header analysis as proposed in [41] was used along with some modifications. In the mentioned paper, the authenticity of the email was checked using a matching technique. If the first Received From field had the same domain as From or localhost or current email account or forwarding email account or Received SPF address then email was marked legitimate else, phishing. As an enhancement, more fields like the CC field, the BCC field and the Message-ID field domains were included while matching the Received From field domain. Another check was added so that if the Received From field domain is different from the ones before and after it, both of which are same, the email is marked as phishing. For example if the email goes from domain a.relay.com to b.relay.com to a.relay.com again, it is suspicious and the email is marked phishing. This is to support the fact that a legitimate email will not contain such cycles where it leaves a domain and then reenters after going through some other domain.

4.6 Phish-IDetector

In Phish-IDetector, as explained in chapter 4, the Message-IDs are extracted from all the emails. Features are extracted using N-gram analysis and the classification is dome using machine learning classifiers in Weka.

4.7 Results Collation

For combining the results from these classifiers we used majority voting. Table 4.1 shows the results for the set of emails for which all 3 classifiers had predictions. Table 4.2 shows the results for the complete datasets. Some of these emails did not have predictions from all 3 classes. For these emails, the final class was decided as the prediction from Phish-IDetector since it had the highest TPR.

Table 4.1: The Collated Results for the Emails classified by all 3 classifiers

All3Present					
Dataset	**Phish**	**Legitimate**	**Total**	**TPR**	**FPR**
EasyHam	5	1511	1516	99.67	0.33
EasyHam1	17	733	750	97.733	2.267
EasyHam2	9	412	421	97.862	2.138
HardHam	6	144	150	96	4
HardHam1	1	74	75	98.667	1.333
EasyHamPhish	1306	27	1333	97.974	2.026
EasyHam1Phish	1315	18	1333	98.65	1.35
EasyHam2Phish	1315	18	1333	98.65	1.35
HardHamPhish	1315	18	1333	98.65	1.35
HardHam1Phish	1315	18	1333	98.645	1.35

Table 4.2: The Collated Results for All Datasets

Dataset	**Phish**	**Legitimate**	**Total**	**TPR**	**FPR**
EasyHam	15	5036	5051	99.703	0.297
EasyHam1	134	2366	2500	94.64	5.36
EasyHam2	55	1345	1400	96.071	3.928
HardHam	7	493	500	98.6	1.4
HardHam1	2	248	250	99.2	0.8
EasyHamPhish	4335	215	4550	95.275	4.725
EasyHam1Phish	4517	33	4550	99.275	0.725
EasyHam2Phish	4528	22	4550	99.516	0.484
HardHamPhish	4529	21	4550	99.538	0.462
HardHam1Phish	4529	21	4550	99.538	0.462

Chapter 5

Header-Domain Analysis

As mentioned in Chapter 1 and 2, each email contains information about the path it has traveled from the sender to the receiver. The domains in the header fields give a good approximation of this path. This header field is a string following a certain basic format. It also contains information which can be used to trace the path of the email. The domains closer to the receiver's side cannot be altered easily and it provides important information about the trail the email has followed. Figure 5.1 shows an example of an email header with the header domains.

5.1 Prediction

Studying the mechanism of email transfer and delivery closely, we find that each email can provide us with information to trace its path. Also, we feel the domain headers [Please refer Chapter 2 for description.] give a good indication of this path.

Our prediction is that the string of all the domain headers in an email would show signs of obfuscation in case of phishing emails where the phisher is tricking users to believe that the email has come from an authentic source. Also, since phishers are distributed across the world, the header domains can indicate the path and hence their location which could be key to distinguishing them from the legitimate senders. Furthermore, it may be that phishers could use source routing for the emails, where they fix the path for the emails in advance using available open smtp servers resulting in same or similar paths for the phishing emails. Once a path or a part of the path in the form of domain names is found to be associated with phishing emails, another email with the same path will most likely be phish as well. To test our hypothesis we chose to combine two of the most popular and effective techniques of phishing email classification: n-gram analysis [25], [22], [39] and machine learning [16], [20], [1].

5.2 The Overall Approach

The main aim being classification of emails as phishing or legitimate, our approach works using n-gram analysis of the string of domain names present in the header of the emails. Using popular machine-learning classifiers proved to be effective for phishing email classification, we obtain good results. The method used for classification was chosen to be 10-fold cross validation due to its known effectiveness and universal acceptance. The advantage of our system is its combination of simplicity and effectiveness. Though minimal information is required and the whole process involves no complex steps, the results are promising. Since we used different datasets

from various sources [described in Section 5.4.] and still got consistent results, it proves the robustness of our method.

Apart from the domains, we derive information about the email's path [Please refer to Chapter 8 for details] by using four different types of analyses.

Finally we conduct several experiments involving n-gram features from the header domains.

5.3 Architecture

Our system consists of the following main components. Figure 5.2 is a diagrammatic representation of our system. The individual components are summarized below.

5.3.1 Domain Extraction Component

This component is responsible for extracting the domains from the email headers. The raw emails with full headers serve as the input. As mentioned, the string of header fields contain information about an email's path. Some of the header fields have this information in the format of an email address like LHS@RHS, and we extract extract only the RHS part to get the domain names. Other fields have just the domain name and we extract the full domains. All the domain names are collected from each email and stored in a single string, separated by commas. The string is the output of this component. For example in the header shown in Figure 5.1, the string will include these domains: citizensbank.com, login.monkey.org, mail1.monkey.org,

funky.monkey.org, mail2.monkey.org, and so on. TLD Removal: We formed new datasets by removing the (Top Level Domains) TLDs from the collected domains to get rid of any bias caused due to the TLD differences between the legitimate and phishing datsets. The results did deteriorate but only a little which shows that the TLD difference had a very small contribution in the classification success. A more detailed analysis is done in section 5.7.

5.3.2 Data File Creation Component

After the domain extraction is done, this component handles the creation of the data file, which consists of the class information for all the emails along with their respective header domains string. For each email we determined the given label based on the dataset, i.e., "phishing" or "legit" (legitimate) class and put the extracted domain headers beside each label. Hence, we create a csv file with two columns. First containing the class label and the second containing a single string of header domains for each email.

5.3.3 N-gram Analysis Component

Next, the n-gram analysis component takes the output csv file from the previous component and performs n-gram analysis on the information. We decided to do n-gram analysis of the collected header domains, as this kind of analysis is able to capture the structure present in any text or string. Also, this method enabled us to represent the data in numeric format acceptable to most classifiers in Weka. This

analysis generates unique n-gram features represented as Unicode code point of the characters. For example, the Unicode code point for "a" is 97 so the 1-gram "a" will be represented as 97. The feature extracted is the frequency of the n-gram in the header domains. Hence, the original data collected was transformed and represented in the arff format. The arff files were also converted to .svm format, the input format for the confidence weighted algorithm [28], using a python script.

5.3.4 Classification Component

Once we obtained and represented the data in arff format, we passed it on to the classification component. Here the following seven classifiers were run on the arff file using Weka 3.6: RF, J48, Bagging, AttSel, SMO, BLR and NBMultinomial. The confidence weighted algorithm mentioned in Chapter 3 was also used for classification.

5.4 Data Sets and Classifiers

We have used two publicly available datasets and two datasets collected from volunteers. Email header domains were collected from these datasets separately. In total we had 3392 phishing emails from a private dataset created by Dr. Jose Nazario [31], 2949 legitimate emails from [14] 197 phishing and 4986 legitimate emails from one of the author's inbox.

We ran experiments on combinations of the phishing emails sets combined with

each of the above mentioned sets of legitimate emails and experimented with both balanced datasets as well as unbalanced datasets to study the effect on the results. For the unbalanced datasets we had some with more phishing emails than legitimate and some vice versa. These datasets are named according to the sets involved in creating the final dataset. The names and description of the datasets are given below.

5.4.1 Unbalanced Datasets

1. CSDMCNPN
2. CSDMCRV
3. RVLNPN
4. RVLRV

5.4.2 Balanced Datasets

1. BalCSDMCNPN (2949 legit and phish emails each)
2. BalCSDMCRV (197 legit and phish emails each)
3. BalRVLNPN (3392 legit and phish emails each)
4. BalRVLRV (197 legit and phish emails each)

We used Weka version 3.6 which is basically a collection of machine learning algorithms for data mining tasks. It was chosen because of its wide acceptability,

popularity and its ease of use. It has previously been used for phishing detection by [20] and [11]. Weka provided us an easy method of comparing the performance of several classifiers on our datasets and choosing the best among them. We ran the experiments with around 7 classifiers and chose the best among them. Each of them is explained here in brief.

Random Forest (RF) classifier [Please refer Chapter 3].

J48 [Please refer Chapter 3].

SMO [Please refer Chapter 3].

Bootstrap Aggregating or **Bagging** [Please refer Chapter 3].

AttributeselectedClassifer This (AttSel) classifier first does attribute selection and reduces the dimensionality of the training and testing sets before running the classifier. It is useful in removing redundant attributes and thus improving classification. We used the default options for this classifier in our experiments.

BayesianLogisticRegression (BLR) is an implementation of bayesian logistic regression for both Gaussian and Laplace priors and more details can be found at [18].

NaiveBayes (NB) [Please refer Chapter 3].

Apart from Weka, we also used an online confidence weighted algorithm for classification [28]. The main advantage of an online learning algorithm is its speed. Being much faster than Weka, we were able to classify even higher order n-gram files using confidence weighted algorithm. Since online algorithms have the capability to learn

from each instance and then discard it immediately, without storing the whole set of instances, it can run much faster than the batch algorithms. Specially in our case, where the feature set increases exponentially for every higher n-gram, it was a great option.

5.5 Experiment Including IPs in Domains

The first set of experiments involved only the domains from the emails and not the IP addresses contained in them. We realized that this would result in loss of data and eventually sub standard results. So we repeated the experiments taking the IPs in the header into account. This did result in the increase in processing time since the extra step of extracting domain from IP was added.

5.6 Results

The results for the various experiments conducted are summarized in this section.

Tables 5.1 to 5.5 summarize the TPR and FPR values of the experiments on the balanced dataset RVLNPN with full domains. Results till 5-grams are reported here. These show that for some classifiers like NBMultinomial the detection improves with higher order n-grams whereas for some classifiers like SMO and RF performance slightly decreases as n-gram order increases.

Figure 5.1: An Email Header With Header Domains

Return-Path: <businessclients.refvi842377742.gps@citizensbank.com>X-Original-To: jose@login.monkey.org
Delivered-To: jose@login.monkey.org
Received: from mail1.monkey.org (mail1.monkey.org [152.160.49.212])
by funky.monkey.org (Postfix) with ESMTP id 77D7546909
for <jose@login.monkey.org>; Thu, 9 Aug 2007 19:12:11 -0400 (EDT)
Received: from mail2.monkey.org (mail2.monkey.org [204.181.64.8])
by mail1.monkey.org (Postfix) with ESMTP id 4823F131FD3B
for <jose@monkey.org>; Thu, 9 Aug 2007 19:12:11 -0400 (EDT)
Received: from 84.94.90.70.cable.012.net.il (84.94.90.70.cable.012.net.il [84.94.90.70])
by mail2.monkey.org (Postfix) with SMTP id 1367E6FA1C4
for <jose@monkey.org>; Thu, 9 Aug 2007 19:12:05 -0400 (EDT)
Received: from servebeer.com [109.181.144.192]
by gayrated.com with SMTP id GICBRU6K6V
for <jose@monkey.org>; Thu, 09 Aug 2007 18:12:01 -0600
Received: from bmla.com (bmla.com.timeanddate.com [130.106.111.24])
by surfjunky.com with SMTP id EK42NELFKN
for <jose@monkey.org>; Thu, 09 Aug 2007 23:08:01 -0100
From: "Citizens Bank and Charter One Bank" businessclients.refvi842377742.gps@citizensbank.com
To: "Jose" jose@monkey.org
Subject: Citizens Bank and Charter One Bank: customer details confirmation! (message id: o85393182q)
X-MimeOLE: Produced By Microsoft MimeOLE V6.00.2800.1165User-Agent: MIME-tools 5.503 (Entity 5.501)
Content-Type: multipart/alternative; boundary="--A7DGA8EL9LENVJZY6“
Message-Id: 20070809231205.1367E6FA1C4@mail2.monkey.org

Table 5.1: 1gramFullDomainsBalRVLNPN

1gramFullDomainsBalRVLNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	0.999	0.001
NBMultinomial	0.997	0.003
SMO	1	0
AttSel	0.998	0.002
Bagging	0.999	0.001
J48	0.999	0.001
RF	1	0

Figure 5.2: Architecture of Header Domain Analysis System

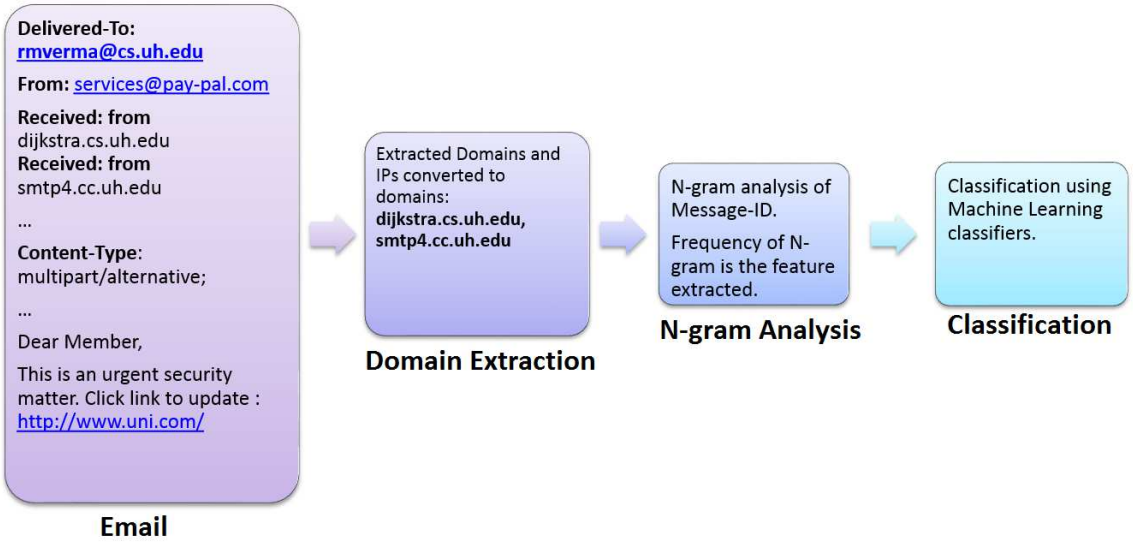


Table 5.2: 2gramFullDomainBalRVLNPN

2gramFullDomainBalRVLNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	1	0
NBMultinomial	0.997	0.003
SMO	1	0
AttSel	0.998	0.002
Bagging	0.999	0.001
J48	0.999	0.001
RF	0.999	0.001

Table 5.3: 3gramFullDomainsBalRVLNPN

3gramFullDomainsBalRVLNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	1	0
NBMultinomial	0.997	0.003
SMO	1	0
AttSel	0.998	0.002
Bagging	0.999	0.001
J48	0.999	0.001
RF	0.999	0.001

Table 5.4: 4gramFullDomainsBalRVLNPN

4gramFullDomainsBalRVLNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	0.999	0.001
NBMultinomial	0.998	0.002
SMO	0.999	0.001
AttSel	0.998	0.002
Bagging	0.999	0.001
J48	0.999	0.001
RF	0.999	0.001

Table 5.5: 5gramFullDomainsBalRVLNPN

5gramFullDomainsBalRVLNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	1	0
NBMultinomial	0.998	0.002
SMO	0.999	0.001
AttSel	0.997	0.003
Bagging	0.999	0.001
J48	0.999	0.001
RF	0.999	0.001

Tables 5.6 to 5.10 summarize the TPR and FPR values of the experiments on the balanced dataset RVLNPN having domains without TLDs. Results till 5-grams are reported here. These show that for some classifiers like NBMultinomial and BLR the detection improves with higher order n-grams whereas for some classifiers like AttSel performance slightly decreases as n-gram order increases.

Tables 5.11 to 5.15 summarize the TPR and FPR values of the experiments on the unbalanced dataset RVLNPN having full domains. Results till 5-grams are reported here. These show that for some classifiers like AttSel the detection improves with higher order n-grams whereas for some classifiers like BLR and SMO performance slightly decreases as n-gram order increases.

Tables 5.16 to 5.20 summarize the TPR and FPR values of the experiments on the unbalanced dataset RVLNPN having domains without TLDs. Results till 5-grams are reported here. These show that for some classifiers like BLR and SMO the detection improves with higher order n-grams whereas performance does not decreases as n-gram order increases for any of the classifiers.

Table 5.6: 1gramNoTLDDomainsBalRVLNPN

1gramNoTLDDomainsBalRVLNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	0.999	0.001
NBMultinomial	0.997	0.003
SMO	0.999	0.001
AttSel	0.998	0.002
Bagging	0.998	0.002
J48	0.999	0.001
RF	0.999	0.001

Table 5.7: 2gramNoTLDDomainsBalRVLNPN

2gramNoTLDDomainsBalRVLNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	1	0
NBMultinomial	0.997	0.003
SMO	0.999	0.001
AttSel	0.997	0.003
Bagging	0.999	0.001
J48	0.999	0.001
RF	0.999	0.001

Table 5.8: 3gramNoTLDDomainsBalRVLNPN

3gramNoTLDDomainsBalRVLNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	1	0
NBMultinomial	0.998	0.002
SMO	0.999	0.001
AttSel	0.997	0.003
Bagging	0.999	0.001
J48	0.999	0.001
RF	0.999	0.001

Table 5.9: 4gramNoTLDDomainsBalRVLNPN

4gramNoTLDDomainsBalRVLNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	1	0
NBMultinomial	0.998	0.002
SMO	0.999	0.001
AttSel	0.997	0.003
Bagging	0.999	0.001
J48	0.999	0.001
RF	0.999	0.001

Table 5.10: 5gramNoTLDDomainsBalRVLNPN

5gramNoTLDDomainsBalRVLNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	1	0
NBMultinomial	0.998	0.002
SMO	0.999	0.001
AttSel	0.997	0.003
Bagging	0.999	0.001
J48	0.999	0.001
RF	0.999	0.001

Table 5.11: 1gramFullDomainsRVLNPN

1gramFullDomainsRVLNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	0.999	0.001
NBMultinomial	0.997	0.003
SMO	1	0
AttSel	0.999	0.001
Bagging	1	0.001
J48	1	0
RF	1	0.001

Table 5.12: 2gramFullDomainsRVLNPN

2gramFullDomainsRVLNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	1	0
NBMultinomial	0.997	0.003
SMO	1	0.001
AttSel	1	0.001
Bagging	0.999	0.001
J48	0.999	0.001
RF	1	0

Table 5.13: 3gramFullDomainsRVLNPN

3gramFullDomainsRVLNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	1	0.001
NBMultinomial	0.998	0.001
SMO	1	0.001
AttSel	1	0.001
Bagging	0.999	0.001
J48	0.999	0.001
RF	1	0.001

Table 5.14: 4gramFullDomainsRVLNPN

4gramFullDomainsRVLNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	1	0.001
NBMultinomial	0.998	0.001
SMO	0.999	0.001
AttSel	1	0.001
Bagging	0.999	0.001
J48	0.999	0.001
RF	0.999	0.001

Table 5.15: 5gramFullDomainsRVLNPN

5gramFullDomainsRVLNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	0.999	0.001
NBMultinomial	0.998	0.001
SMO	0.999	0.001
AttSel	1	0.001
Bagging	0.999	0.001
J48	0.999	0.001
RF	0.999	0.001

Table 5.16: 1gramNoTLDDomainsRVLNPN

1gramNoTLDDomainsRVLNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	0.999	0.001
NBMultinomial	0.996	0.004
SMO	0.999	0.001
AttSel	0.998	0.002
Bagging	0.999	0.001
J48	0.999	0.001
RF	1	0

Table 5.17: 2gramNoTLDDomainsRVLNPN

2gramNoTLDDomainsRVLNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	1	0
NBMultinomial	0.997	0.003
SMO	1	0
AttSel	0.999	0.001
Bagging	0.999	0.001
J48	0.999	0
RF	1	0

Table 5.18: 3gramNoTLDDomainsRVLNPN

3gramNoTLDDomainsRVLNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	1	0
NBMultinomial	0.998	0.003
SMO	1	0
AttSel	0.999	0.001
Bagging	0.999	0
J48	0.999	0
RF	1	0

Table 5.19: 4gramNoTLDDomainsRVLNPN

4gramNoTLDDomainsRVLNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	1	0
NBMultinomial	0.998	0.003
SMO	1	0
AttSel	0.999	0.001
Bagging	0.999	0
J48	0.999	0
RF	1	0

Table 5.20: 5gramNoTLDDomainsRVLNPN

5gramNoTLDDomainsRVLNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	1	0
NBMultinomial	0.998	0.003
SMO	1	0
AttSel	0.999	0.001
Bagging	0.999	0.001
J48	0.999	0
RF	1	0

Table 5.21: 1gramFullDomainsBalCSDMCNPN

1gramFullDomainsBalCSDMCNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	0.998	0.002
NBMultinomial	0.987	0.013
SMO	0.998	0.002
AttSel	0.997	0.003
Bagging	0.999	0.001
J48	0.994	0.006
RF	0.999	0.001

Table 5.22: 2gramFullDomainsBalCSDMCNPN

2gramFullDomainsBalCSDMCNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	0.999	0.001
NBMultinomial	0.996	0.004
SMO	0.999	0.001
AttSel	0.998	0.002
Bagging	0.997	0.003
J48	0.999	0.001
RF	0.999	0.001

Table 5.23: 3gramFullDomainsBalCSDMCNPN

3gramFullDomainsBalCSDMCNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	0.999	0.001
NBMultinomial	0.997	0.003
SMO	0.999	0.001
AttSel	0.998	0.002
Bagging	0.998	0.002
J48	0.999	0.001
RF	0.999	0.001

Tables 5.21 to 5.25 summarize the TPR and FPR values of the experiments on the balanced dataset CSDMCNPN having full domains. Results till 5-grams are reported here. These show that for some classifiers like BLR and SMO the detection improves with higher order n-grams whereas performance does not decreases as n-gram order increases for any of the classifiers.

Tables 5.26 to 5.30 summarize the TPR and FPR values of the experiments on the balanced dataset CSDMCNPN having domains with no TLD. Results till 5-grams are reported here. These show that for some classifiers like BLR and NBMultinomial the detection improves with higher order n-grams whereas performance does not decreases as n-gram order increases for any of the classifiers.

Tables 5.31 to 3.35 summarize the TPR and FPR values of the experiments on the unbalanced dataset CSDMCNPN having full domains. Results till 5-grams are reported here. These show that for some classifiers like BLR and SMO the detection improves with higher order n-grams whereas performance does not decreases as n-gram order increases for any of the classifiers.

Tables 5.36 to 5.40 summarize the TPR and FPR values of the experiments on the balanced dataset CSDMCNPN having domains with no TLD. Results till 5-grams are reported here. These show that for some classifiers like BLR and SMO the detection improves with higher order n-grams whereas performance does not decreases as n-gram order increases for any of the classifiers.

Table 5.24: 4gramFullDomainsBalCSDMCNPN

4gramFullDomainsBalCSDMCNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	0.999	0.001
NBMultinomial	0.997	0.003
SMO	0.999	0.001
AttSel	0.998	0.002
Bagging	0.998	0.002
J48	0.999	0.001
RF	0.999	0.001

Table 5.25: 5gramFullDomainsBalCSDMCNPN

5gramFullDomainsBalCSDMCNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	0.999	0.001
NBMultinomial	0.997	0.003
SMO	0.999	0.001
AttSel	0.998	0.002
Bagging	0.998	0.002
J48	0.999	0.001
RF	0.999	0.001

Table 5.26: 1gramNoTLDDomainsBalCSDMCNPN

1gramNoTLDDomainsBalCSDMCNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	0.998	0.002
NBMultinomial	0.991	0.009
SMO	0.999	0.001
AttSel	0.987	0.013
Bagging	0.998	0.002
J48	0.995	0.005
RF	0.999	0.001

Table 5.27: 2gramNoTLDDomainsBalCSDMCNPN

2gramNoTLDDomainsBalCSDMCNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	0.999	0.001
NBMultinomial	0.997	0.003
SMO	0.999	0.001
AttSel	0.998	0.002
Bagging	0.999	0.001
J48	0.999	0.001
RF	0.999	0.001

Table 5.28: 3gramNoTLDDomainsBalCSDMCNPN

3gramNoTLDDomainsBalCSDMCNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	0.999	0.001
NBMultinomial	0.997	0.003
SMO	0.999	0.001
AttSel	0.998	0.002
Bagging	0.998	0.002
J48	0.999	0.001
RF	0.999	0.001

Table 5.29: 4gramNoTLDDomainsBalCSDMCNPN

4gramNoTLDDomainsBalCSDMCNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	0.999	0.001
NBMultinomial	0.997	0.003
SMO	0.999	0.001
AttSel	0.998	0.002
Bagging	0.998	0.002
J48	0.999	0.001
RF	0.999	0.001

Table 5.30: 5gramNoTLDDomainsBalCSDMCNPN

5gramNoTLDDomainsBalCSDMCNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	0.999	0.001
NBMultinomial	0.997	0.003
SMO	0.999	0.001
AttSel	0.998	0.002
Bagging	0.998	0.002
J48	0.999	0.001
RF	0.999	0.001

Table 5.31: 1gramFullDomainsCSDMCNPN

1gramFullDomainsCSDMCNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	0.998	0.001
NBMultinomial	0.987	0.012
SMO	0.998	0.001
AttSel	0.994	0.007
Bagging	0.998	0.001
J48	0.995	0.005
RF	0.999	0.001

Table 5.32: 2gramFullDomainsCSDMCNPN

2gramFullDomainsCSDMCNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	0.999	0.001
NBMultinomial	0.998	0.002
SMO	0.999	0.001
AttSel	0.997	0.003
Bagging	0.998	0.002
J48	0.998	0.002
RF	0.999	0.001

Table 5.33: 3gramFullDomainsCSDMCNPN

3gramFullDomainsCSDMCNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	0.999	0.001
NBMultinomial	0.997	0.003
SMO	0.999	0.001
AttSel	0.997	0.003
Bagging	0.999	0.001
J48	0.998	0.002
RF	0.999	0.001

Table 5.34: 4gramFullDomainsCSDMCNPN

4gramFullDomainsCSDMCNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	0.999	0.001
NBMultinomial	0.997	0.003
SMO	0.999	0.001
AttSel	0.997	0.003
Bagging	0.999	0.001
J48	0.998	0.002
RF	0.999	0.001

Table 5.35: 5gramFullDomainsCSDMCNPN

5gramFullDomainsCSDMCNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	0.999	0.001
NBMultinomial	0.997	0.003
SMO	0.999	0.001
AttSel	0.997	0.003
Bagging	0.999	0.001
J48	0.998	0.002
RF	0.999	0.001

Table 5.36: 1gramNoTLDDomainsCSDMCNPN

1gramNoTLDDomainsCSDMCNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	0.998	0.002
NBMultinomial	0.986	0.006
SMO	0.999	0
AttSel	0.99	0.019
Bagging	0.999	0.002
J48	0.998	0.003
RF	1	0

Table 5.37: 2gramNoTLDDomainsCSDMCNPN

2gramNoTLDDomainsCSDMCNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	1	0
NBMultinomial	0.997	0.002
SMO	1	0
AttSel	0.995	0.008
Bagging	0.999	0.001
J48	0.998	0.002
RF	1	0

Table 5.38: 3gramNoTLDDomainsCSDMCNPN

3gramNoTLDDomainsCSDMCNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	1	0
NBMultinomial	0.997	0.002
SMO	1	0
AttSel	0.997	0.007
Bagging	0.998	0.002
J48	0.998	0.002
RF	1	0

Table 5.39: 4gramNoTLDDomainsCSDMCNPN

4gramNoTLDDomainsCSDMCNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	1	0
NBMultinomial	0.997	0.002
SMO	1	0
AttSel	0.997	0.007
Bagging	0.998	0.002
J48	0.998	0.002
RF	1	0

Table 5.40: 5gramNoTLDDomainsCSDMCNPN

5gramNoTLDDomainsCSDMCNPN		
Classifier	**TP Rate**	**FP Rate**
BLR	1	0
NBMultinomial	0.996	0.002
SMO	1	0
AttSel	0.997	0.007
Bagging	0.998	0.002
J48	0.998	0.002
RF	1	0

Tables 5.41 to 5.45 summarize the TPR and FPR values of the experiments on the balanced dataset RVLRV having full domains. Results till 5-grams are reported here. These show very poor performance and the reason could be that all the emails are from the same individuals account and for almost all of them the source as well as the destination domains are the same.

Tables 5.46 to 5.50 summarize the TPR and FPR values of the experiments on the balanced dataset RVLRV having domains with no TLD. Results till 5-grams are reported here. These show very poor performance and the reason could be that all the emails are from the same individuals account and for almost all of them the source as well as the destination domains are the same.

Tables 5.51 to 5.55 summarize the TPR and FPR values of the experiments on the unbalanced dataset RVLRV having full domains. Results till 5-grams are reported here. These show much better results than the balanced dataset because the distribution of legitimate and phishing emails are very skewed with only 197 phishing emails vs. 4986 Legitimate ones.

Table 5.41: 1gramFullDomainsBalRVLRV

1gramFullDomainsBalRVLRV		
Classifier	**TP Rate**	**FP Rate**
BLR	0.508	0.492
NBMultinomial	0.5	0.5
SMO	0.495	0.505
AttSel	0.492	0.508
Bagging	0.508	0.492
J48	0.492	0.508
RF	0.503	0.497

Table 5.42: 2gramFullDomainsBalRVLRV

2gramFullDomainsBalRVLRV		
Classifier	**TP Rate**	**FP Rate**
BLR	0.505	0.495
NBMultinomial	0.5	0.5
SMO	0.503	0.497
AttSel	0.492	0.508
Bagging	0.503	0.497
J48	0.492	0.508
RF	0.495	0.505

Table 5.43: 3gramFullDomainsBalRVLRV

3gramFullDomainsBalRVLRV		
Classifier	**TP Rate**	**FP Rate**
BLR	0.505	0.495
NBMultinomial	0.5	0.5
SMO	0.5	0.5
AttSel	0.492	0.508
Bagging	0.503	0.497
J48	0.492	0.508
RF	0.503	0.497

Table 5.44: 4gramFullDomainsBalRVLRV

4gramFullDomainsBalRVLRV		
Classifier	**TP Rate**	**FP Rate**
BLR	0.505	0.495
NBMultinomial	0.505	0.495
SMO	0.503	0.497
AttSel	0.492	0.508
Bagging	0.503	0.497
J48	0.492	0.508
RF	0.503	0.497

Table 5.45: 5gramFullDomainsBalRVLRV

5gramFullDomainsBalRVLRV		
Classifier	**TP Rate**	**FP Rate**
BLR	0.505	0.495
NBMultinomial	0.505	0.495
SMO	0.508	0.492
AttSel	0.492	0.508
Bagging	0.503	0.497
J48	0.492	0.508
RF	0.497	0.503

Table 5.46: 1gramNoTLDDomainsBalRVLRV

1gramNoTLDDomainsBalRVLRV		
Classifier	**TP Rate**	**FP Rate**
BLR	0.497	0.503
NBMultinomial	0.505	0.495
SMO	0.5	0.5
AttSel	0.492	0.508
Bagging	0.5	0.5
J48	0.492	0.508
RF	0.497	0.503

Table 5.47: 2gramNoTLDDomainsBalRVLRV

2gramNoTLDDomainsBalRVLRV		
Classifier	**TP Rate**	**FP Rate**
BLR	0.503	0.497
NBMultinomial	0.505	0.495
SMO	0.5	0.5
AttSel	0.492	0.508
Bagging	0.5	0.5
J48	0.492	0.508
RF	0.497	0.503

Table 5.48: 3gramNoTLDDomainsBalRVLRV

3gramNoTLDDomainsBalRVLRV		
Classifier	**TP Rate**	**FP Rate**
BLR	0.503	0.497
NBMultinomial	0.505	0.495
SMO	0.5	0.5
AttSel	0.492	0.508
Bagging	0.5	0.5
J48	0.492	0.508
RF	0.497	0.503

Table 5.49: 4gramNoTLDDomainsBalRVLRV

4gramNoTLDDomainsBalRVLRV		
Classifier	**TP Rate**	**FP Rate**
BLR	0.5	0.5
NBMultinomial	0.505	0.495
SMO	0.505	0.495
AttSel	0.492	0.508
Bagging	0.5	0.5
J48	0.492	0.508
RF	0.495	0.505

Table 5.50: 5gramNoTLDDomainsBalRVLRV

5gramNoTLDDomainsBalRVLRV		
Classifier	**TP Rate**	**FP Rate**
BLR	0.5	0.5
NBMultinomial	0.505	0.495
SMO	0.495	0.505
AttSel	0.492	0.508
Bagging	0.5	0.5
J48	0.492	0.508
RF	0.495	0.505

Table 5.51: 1gramFullDomainsRVLRV

1gramFullDomainsRVLRV		
Classifier	**TP Rate**	**FP Rate**
BLR	0.962	0.962
NBMultinomial	0.958	0.962
SMO	0.962	0.962
AttSel	0.962	0.962
Bagging	0.962	0.962
J48	0.962	0.962
RF	0.962	0.962

Table 5.52: 2gramFullDomainsRVLRV

2gramFullDomainsRVLRV		
Classifier	**TP Rate**	**FP Rate**
BLR	0.962	0.962
NBMultinomial	0.954	0.962
SMO	0.962	0.962
AttSel	0.962	0.962
Bagging	0.962	0.962
J48	0.962	0.962
RF	0.962	0.962

Table 5.53: 3gramFullDomainsRVLRV

3gramFullDomainsRVLRV		
Classifier	**TP Rate**	**FP Rate**
BLR	0.962	0.962
NBMultinomial	0.954	0.962
SMO	0.962	0.962
AttSel	0.962	0.962
Bagging	0.962	0.962
J48	0.962	0.962
RF	0.962	0.962

Table 5.54: 4gramFullDomainsRVLRV

4gramFullDomainsRVLRV		
Classifier	**TP Rate**	**FP Rate**
BLR	0.962	0.962
NBMultinomial	0.954	0.962
SMO	0.962	0.962
AttSel	0.962	0.962
Bagging	0.962	0.962
J48	0.962	0.962
RF	0.962	0.962

Table 5.55: 5gramFullDomainsRVLRV

5gramFullDomainsRVLRV		
Classifier	**TP Rate**	**FP Rate**
BLR	0.962	0.962
NBMultinomial	0.954	0.962
SMO	0.962	0.962
AttSel	0.962	0.962
Bagging	0.962	0.962
J48	0.962	0.962
RF	0.962	0.962

Table 5.56: 1gramNoTLDDomainsRVLRV

1gramNoTLDDomainsRVLRV		
Classifier	**TP Rate**	**FP Rate**
BLR	0.962	0.962
NBMultinomial	0.959	0.962
SMO	0.962	0.962
AttSel	0.962	0.962
Bagging	0.962	0.962
J48	0.962	0.962
RF	0.962	0.962

Table 5.57: 2gramNoTLDDomainsRVLRV

2gramNoTLDDomainsRVLRV		
Classifier	**TP Rate**	**FP Rate**
BLR	0.962	0.962
NBMultinomial	0.959	0.962
SMO	0.962	0.962
AttSel	0.962	0.962
Bagging	0.962	0.962
J48	0.962	0.962
RF	0.962	0.962

Tables 5.56 to 5.60 summarize the TPR and FPR values of the experiments on the unbalanced dataset RVLRV having domains with no TLD. Results till 5-grams are reported here. These show much better results than the balanced dataset because the distribution of legitimate and phishing emails are very skewed with only 197 phishing emails vs 4986 Legitimate ones.

Tables 5.61 to 5.65 summarize the TPR and FPR values of the experiments on the unbalanced dataset CSDMCRV having full domains. Results till 5-grams are reported here. These results show that the detection rates are fairly constant for all classifiers.

Tables 5.66 to 5.70 summarize the TPR and FPR values of the experiments on the unbalanced dataset CSDMCRV having domains with no TLD. Results till 5-grams are reported here. These results show that the detection rates are fairly constant for all classifiers.

Table 5.58: 3gramNoTLDDomainsRVLRV

3gramNoTLDDomainsRVLRV		
Classifier	**TP Rate**	**FP Rate**
BLR	0.962	0.962
NBMultinomial	0.959	0.962
SMO	0.962	0.962
AttSel	0.962	0.962
Bagging	0.962	0.962
J48	0.962	0.962
RF	0.962	0.962

Table 5.59: 4gramNoTLDDomainsRVLRV

4gramNoTLDDomainsRVLRV		
Classifier	**TP Rate**	**FP Rate**
BLR	0.962	0.962
NBMultinomial	0.959	0.962
SMO	0.962	0.962
AttSel	0.962	0.962
Bagging	0.962	0.962
J48	0.962	0.962
RF	0.962	0.962

Table 5.60: 5gramNoTLDDomainsRVLRV

5gramNoTLDDomainsRVLRV		
Classifier	**TP Rate**	**FP Rate**
BLR	0.962	0.962
NBMultinomial	0.959	0.962
SMO	0.962	0.962
AttSel	0.962	0.962
Bagging	0.962	0.962
J48	0.962	0.962
RF	0.962	0.962

Table 5.61: 1gramFullDomainsBalCSDMCRV

1gramFullDomainsBalCSDMCRV		
Classifier	**TP Rate**	**FP Rate**
BLR	1	0
NBMultinomial	1	0
SMO	0.997	0.003
AttSel	0.995	0.005
Bagging	0.995	0.005
J48	0.995	0.005
RF	0.997	0.003

Table 5.62: 2gramFullDomainsBalCSDMCRV

2gramFullDomainsBalCSDMCRV		
Classifier	**TP Rate**	**FP Rate**
BLR	1	0
NBMultinomial	1	0
SMO	1	0
AttSel	0.995	0.005
Bagging	0.995	0.005
J48	0.995	0.005
RF	0.997	0.003

Table 5.63: 3gramFullDomainsBalCSDMCRV

3gramFullDomainsBalCSDMCRV		
Classifier	**TP Rate**	**FP Rate**
BLR	1	0
NBMultinomial	1	0
SMO	0.997	0.003
AttSel	0.995	0.005
Bagging	0.995	0.005
J48	0.995	0.005
RF	0.997	0.003

Table 5.64: 4gramFullDomainsBalCSDMCRV

4gramFullDomainsBalCSDMCRV		
Classifier	**TP Rate**	**FP Rate**
BLR	1	0
NBMultinomial	1	0
SMO	0.997	0.003
AttSel	0.995	0.005
Bagging	0.995	0.005
J48	0.995	0.005
RF	0.997	0.003

Table 5.65: 5gramFullDomainsBalCSDMCRV

5gramFullDomainsBalCSDMCRV		
Classifier	**TP Rate**	**FP Rate**
BLR	1	0
NBMultinomial	1	0
SMO	0.997	0.003
AttSel	0.995	0.005
Bagging	0.995	0.005
J48	0.995	0.005
RF	0.997	0.003

Table 5.66: 1gramNoTLDBalCSDMCRV

1gramNoTLDBalCSDMCRV		
Classifier	**TP Rate**	**FP Rate**
BLR	1	0
NBMultinomial	1	0
SMO	0.995	0.005
AttSel	0.995	0.005
Bagging	0.995	0.005
J48	0.995	0.005
RF	0.997	0.003

Table 5.67: 2gramNoTLDBalCSDMCRV

2gramNoTLDBalCSDMCRV		
Classifier	**TP Rate**	**FP Rate**
BLR	1	0
NBMultinomial	1	0
SMO	0.997	0.003
AttSel	0.992	0.008
Bagging	0.992	0.008
J48	0.992	0.008
RF	0.997	0.003

Table 5.68: 3gramNoTLDBalCSDMCRV

3gramNoTLDBalCSDMCRV		
Classifier	**TP Rate**	**FP Rate**
BLR	1	0
NBMultinomial	1	0
SMO	0.997	0.003
AttSel	0.995	0.005
Bagging	0.995	0.005
J48	0.995	0.005
RF	0.997	0.003

Table 5.69: 4gramNoTLDBalCSDMCRV

4gramNoTLDBalCSDMCRV		
Classifier	**TP Rate**	**FP Rate**
BLR	1	0
NBMultinomial	1	0
SMO	0.997	0.003
AttSel	0.995	0.005
Bagging	0.995	0.005
J48	0.995	0.005
RF	0.997	0.003

Table 5.70: 5gramNoTLDBalCSDMCRV

5gramNoTLDBalCSDMCRV		
Classifier	**TP Rate**	**FP Rate**
BLR	1	0
NBMultinomial	1	0
SMO	0.997	0.003
AttSel	0.995	0.005
Bagging	0.995	0.005
J48	0.995	0.005
RF	0.997	0.003

Table 5.71: 1gramFullDomainsCSDMCRV

1gramFullDomainsCSDMCRV		
Classifier	**TP Rate**	**FP Rate**
BLR	1	0
NBMultinomial	1	0
SMO	1	0.005
AttSel	1	0.005
Bagging	1	0.005
J48	1	0.005
RF	1	0.005

Table 5.72: 2gramFullDomainsCSDMCRV

2gramFullDomainsCSDMCRV		
Classifier	**TP Rate**	**FP Rate**
BLR	1	0
NBMultinomial	0.999	0
SMO	1	0.005
AttSel	1	0.005
Bagging	1	0.005
J48	1	0.005
RF	1	0.005

Table 5.73: 3gramFullDomainsCSDMCRV

3gramFullDomainsCSDMCRV		
Classifier	**TP Rate**	**FP Rate**
BLR	1	0
NBMultinomial	0.998	0
SMO	1	0.005
AttSel	1	0.005
Bagging	1	0.005
J48	1	0.005
RF	1	0.005

Table 5.74: 4gramFullDomainsCSDMCRV

4gramFullDomainsCSDMCRV		
Classifier	**TP Rate**	**FP Rate**
BLR	1	0
NBMultinomial	0.994	0
SMO	1	0.005
AttSel	1	0.005
Bagging	1	0.005
J48	1	0.005
RF	1	0.005

Tables 5.71 to 5.75 summarize the TPR and FPR values of the experiments on the unbalanced dataset CSDMCRV having full domains. Results till 5-grams are reported here. These results show that the detection rates are fairly constant for all classifiers except NBMultinomial.

Tables 5.76 to 5.80 summarize the TPR and FPR values of the experiments on the unbalanced dataset CSDMCRV having domains with no TLD. Results till 5-grams are reported here. These results show that the detection rates are fairly constant for all classifiers except NBMultinomial.

Tables 5.81 and 5.82 summarize the TPR and FPR values of the Confidence-Weighted algorithm experiments on the balanced and unbalanced dataset CSDMC-NPN having domains with no TLD. Results till 10-grams are reported here. These results show that the detection rates are fairly constant for all n-grams.

Tables 5.83 and 5.84 summarize the TPR and FPR values of the Confidence-Weighted algorithm experiments on the balanced and unbalanced dataset CSDM-CNPN having full domains. Results till 10-grams are reported here. These results show that the detection rates are fairly constant for all n-grams.

Tables 5.85 and 5.86 summarize the TPR and FPR values of the Confidence-Weighted algorithm experiments on the balanced and unbalanced dataset RVLNPN having full domains. Results till 10-grams are reported here. These results show that the detection rates are fairly constant for all n-grams.

Tables 5.87 and 5.88 summarize the TPR and FPR values of the Confidence-Weighted algorithm experiments on the balanced and unbalanced dataset RVLNPN

having domains with no TLD. Results till 10-grams are reported here. These results show that the detection rates are fairly constant for all n-grams.

Table 5.75: 5gramFullDomainsCSDMCRV

5gramFullDomainsCSDMCRV		
Classifier	**TP Rate**	**FP Rate**
BLR	1	0
NBMultinomial	0.988	0.001
SMO	1	0.005
AttSel	1	0.005
Bagging	1	0.005
J48	1	0.005
RF	1	0.005

Table 5.76: 1gramNoTLDCSDMCRV

1gramNoTLDCSDMCRV		
Classifier	**TP Rate**	**FP Rate**
BLR	1	0
NBMultinomial	1	0
SMO	1	0.005
AttSel	1	0.005
Bagging	1	0.005
J48	1	0.005
RF	1	0.005

Table 5.77: 2gramNoTLDCSDMCRV

2gramNoTLDCSDMCRV		
Classifier	**TP Rate**	**FP Rate**
BLR	1	0
NBMultinomial	0.999	0
SMO	1	0.005
AttSel	1	0.005
Bagging	1	0.005
J48	1	0.005
RF	1	0.005

Table 5.78: 3gramNoTLDCSDMCRV

3gramNoTLDCSDMCRV		
Classifier	**TP Rate**	**FP Rate**
BLR	1	0
NBMultinomial	0.994	0
SMO	1	0.005
AttSel	1	0.005
Bagging	1	0.005
J48	1	0.005
RF	1	0.005

Table 5.79: 4gramNoTLDCSDMCRV

4gramNoTLDCSDMCRV		
Classifier	**TP Rate**	**FP Rate**
BLR	1	0
NBMultinomial	0.988	0.001
SMO	1	0.005
AttSel	1	0.005
Bagging	1	0.005
J48	1	0.005
RF	1	0.005

Table 5.80: 5gramNoTLDCSDMCRV

5gramNoTLDCSDMCRV		
Classifier	**TP Rate**	**FP Rate**
BLR	1	0
NBMultinomial	0.983	0.001
SMO	1	0.005
AttSel	1	0.005
Bagging	1	0.005
J48	1	0.005
RF	1	0.005

Table 5.81: Confidence-Weighted Results for NoTLDDomainsBalCSDMCNPN

NoTLDDomainsBalCSDMCNPN		
Gram	**TP Rate**	**FP Rate**
1	0.997965412	0.00169549
2	1	0.00169549
3	1	0.00169549
4	1	0.00169549
5	1	0.00169549
6	1	0.00169549
7	1	0.00169549
8	1	0.00169549
9	1	0.00169549
10	1	0.00169549

Table 5.82: Confidence-Weighted Results for NoTLDDomainsCSDMCNPN

NoTLDDomainsCSDMCNPN		
Gram	**TP Rate**	**FP Rate**
1	0.998643608	0.000147406
2	1	0
3	1	0
4	1	0
5	1	0
6	1	0.000294811
7	1	0.000294811
8	1	0.000294811
9	1	0
10	1	0

Table 5.83: Confidence-Weighted Results for FullDomainsBalCSDMCNPN

FullDomainsBalCSDMCNPN		
Gram	**TP Rate**	**FP Rate**
1	0.999660902	0.001356392
2	0.799392097	0.199523323
3	1	0.001356392
4	1	0.001356392
5	1	0.001356392
6	1	0.001356392
7	1	0.001356392
8	1	0.001356392
9	1	0.001356392
10	1	0.001356392

Table 5.84: Confidence-Weighted Results for FullDomainsCSDMCNPN

FullDomainsCSDMCNPN		
Gram	**TP Rate**	**FP Rate**
1	0.99830451	0.000147406
2	1	0
3	1	0
4	1	0
5	1	0
6	1	0
7	1	0.000294811
8	1	0.000294811
9	1	0
10	1	0

Table 5.85: Confidence-Weighted Results for FullDomainsBalRVLNPN

FullDomainsBalRVLNPN		
Gram	**TP Rate**	**FP Rate**
1	0.999115566	0.000589623
2	1	0.001356392
3	0.598996656	0.399587345
4	1	0.001356392
5	1	0.001356392
6	1	0.001356392
7	1	0.001356392
8	1	0.001356392
9	1	0.001356392
10	1	0.001356392

Table 5.86: Confidence-Weighted Results for FullDomainsRVLNPN

FullDomainsRVLNPN		
Gram	**TP Rate**	**FP Rate**
1	0.999115566	0.000589623
2	1	0.001356392
3	0.598996656	0.399587345
4	1	0.001356392
5	1	0.001356392
6	1	0.001356392
7	1	0.001356392
8	1	0.001356392
9	1	0.001356392
10	1	0.001356392

Table 5.87: Confidence-Weighted Results for NoTLDDomainsBalRVLNPN

NoTLDDomainsBalRVLNPN		
Gram	**TP Rate**	**FP Rate**
1	1	0.000589623
2	1	0.000294811
3	0.999410377	0.000884434
4	1	0.000589623
5	0.999410377	0.000589623
6	1	0.000294811
7	1	0.000589623
8	1	0.000589623
9	1	0.000589623
10	0.799822852	0.200765381

Table 5.88: Confidence-Weighted Results for NoTLDDomainsRVLNPN

NoTLDDomainsRVLNPN		
Gram	**TP Rate**	**FP Rate**
1	0.999598877	0
2	1	0
3	1	0
4	0.999598877	0.000589623
5	1	0.000294811
6	1	0
7	0.631010265	0.238378639
8	1	0
9	0.6332907	0.241702398
10	1	0

Table 5.89: Results for 1-gram features of the full domains dataset

1gramFullDomainsFromHeader&IPsNoMsgIdCSDMCRVL+NPNRV							
Class	**Classifier**	**TPR**	**FPR**	**Precision**	**Recall**	**FMeasure**	**ROC**
legit	**ABoost**	0.996	0.004	0.998	0.996	0.997	1
phish	**ABoost**	0.996	0.004	0.991	0.996	0.994	1
Wghtd	**ABoost**	0.996	0.004	0.996	0.996	0.996	1
legit	**AttSel**	0.998	0.074	0.968	0.998	0.983	0.962
phish	**AttSel**	0.926	0.002	0.995	0.926	0.959	0.962
Wghtd	**AttSel**	0.976	0.051	0.976	0.976	0.975	0.962
legit	**Bagging**	1	0.001	0.999	1	1	1
phish	**Bagging**	0.999	0	1	0.999	0.999	1
Wghtd	**Bagging**	1	0.001	1	1	1	1
legit	**BLR**	0.997	0.059	0.974	0.997	0.985	0.969
phish	**BLR**	0.941	0.003	0.993	0.941	0.967	0.969
Wghtd	**BLR**	0.98	0.041	0.98	0.98	0.98	0.969
legit	**J48**	0.998	0.001	0.999	0.998	0.999	0.999
phish	**J48**	0.999	0.002	0.995	0.999	0.997	0.999
Wghtd	**J48**	0.998	0.001	0.998	0.998	0.998	0.999
legit	**NB**	0.998	0.074	0.968	0.998	0.983	0.967
phish	**NB**	0.926	0.002	0.995	0.926	0.959	0.991
Wghtd	**NB**	0.976	0.051	0.976	0.976	0.975	0.975
legit	**RF**	1	0.001	1	1	1	1
phish	**RF**	0.999	0	1	0.999	1	1
Wghtd	**RF**	1	0	1	1	1	1
legit	**SMO**	1	0	1	1	1	1
phish	**SMO**	1	0	1	1	1	1
Wghtd	**SMO**	1	0	1	1	1	1

Table 5.90: Results for 2-gram features of the full domains dataset

2gramFullDomainsFromHeader&IPsNoMsgIdCSDMCRVL+NPNRV							
Class	**Classifier**	**TPR**	**FPR**	**Precision**	**Recall**	**FMeasure**	**ROC**
legit	**ABoost**	0.999	0.004	0.998	0.999	0.999	1
phish	**ABoost**	0.996	0.001	0.998	0.996	0.997	1
Wghtd	**ABoost**	0.998	0.003	0.998	0.998	0.998	1
legit	**AttSel**	1	0.004	0.998	1	0.999	0.998
phish	**AttSel**	0.996	0	0.999	0.996	0.998	0.998
Wghtd	**AttSel**	0.999	0.003	0.999	0.999	0.999	0.998
legit	**Bagging**	1	0.001	1	1	1	1
phish	**Bagging**	0.999	0	1	0.999	0.999	1
Wghtd	**Bagging**	1	0.001	1	1	1	1
legit	**BLR**	1	0	1	1	1	1
phish	**BLR**	1	0	1	1	1	1
Wghtd	**BLR**	1	0	1	1	1	1
legit	**J48**	1	0.002	0.999	1	1	0.999
phish	**J48**	0.998	0	1	0.998	0.999	0.999
Wghtd	**J48**	0.999	0.001	0.999	0.999	0.999	0.999
legit	**NB**	1	0.074	0.968	1	0.983	0.97
phish	**NB**	0.926	0	0.999	0.926	0.961	0.98
Wghtd	**NB**	0.977	0.051	0.978	0.977	0.977	0.974
legit	**RF**	1	0.001	1	1	1	1
phish	**RF**	0.999	0	1	0.999	0.999	1
Wghtd	**RF**	1	0.001	1	1	1	1
legit	**SMO**	1	0	1	1	1	1
phish	**SMO**	1	0	1	1	1	1
Wghtd	**SMO**	1	0	1	1	1	1

Table 5.91: Results for 3-gram features of the full domains dataset

3gramFullDomainsFromHeader&IPsNoMsgIdCSDMCRVL+NPNRV							
Class	**Classifier**	**TPR**	**FPR**	**Precision**	**Recall**	**FMeasure**	**ROC**
legit	**ABoost**	0.999	0.004	0.998	0.999	0.999	1
phish	**ABoost**	0.996	0.001	0.998	0.996	0.997	1
Wghtd	**ABoost**	0.998	0.003	0.998	0.998	0.998	1
legit	**AttSel**	1	0.004	0.998	1	0.999	0.998
phish	**AttSel**	0.996	0	0.999	0.996	0.998	0.998
Wghtd	**AttSel**	0.999	0.003	0.999	0.999	0.999	0.998
legit	**Bagging**	1	0.001	1	1	1	1
phish	**Bagging**	0.999	0	1	0.999	0.999	1
Wghtd	**Bagging**	1	0.001	1	1	1	1
legit	**BLR**	1	0	1	1	1	1
phish	**BLR**	1	0	1	1	1	1
Wghtd	**BLR**	1	0	1	1	1	1
legit	**J48**	1	0.002	0.999	1	1	0.999
phish	**J48**	0.998	0	1	0.998	0.999	0.999
Wghtd	**J48**	0.999	0.001	0.999	0.999	0.999	0.999
legit	**NB**	0.999	0.074	0.968	0.999	0.983	0.971
phish	**NB**	0.926	0.001	0.999	0.926	0.961	0.98
Wghtd	**NB**	0.977	0.051	0.977	0.977	0.976	0.974
legit	**RF**	1	0	1	1	1	1
phish	**RF**	1	0	1	1	1	1
Wghtd	**RF**	1	0	1	1	1	1
legit	**SMO**	1	0	1	1	1	1
phish	**SMO**	1	0	1	1	1	1
Wghtd	**SMO**	1	0	1	1	1	1

Table 5.92: Results for 4-gram features of the full domains dataset

4gramFullDomainsFromHeader&IPsNoMsgIdCSDMCRVL+NPNRV							
Class	**Classifier**	**TPR**	**FPR**	**Precision**	**Recall**	**FMeasure**	**ROC**
legit	**ABoost**	0.999	0.004	0.998	0.999	0.999	1
phish	**ABoost**	0.996	0.001	0.998	0.996	0.997	1
Wghtd	**ABoost**	0.998	0.003	0.998	0.998	0.998	1
legit	**AttSel**	1	0.004	0.998	1	0.999	0.998
phish	**AttSel**	0.996	0	0.999	0.996	0.998	0.998
Wghtd	**AttSel**	0.999	0.003	0.999	0.999	0.999	0.998
legit	**Bagging**	1	0.001	1	1	1	1
phish	**Bagging**	0.999	0	1	0.999	0.999	1
Wghtd	**Bagging**	1	0.001	1	1	1	1
legit	**BLR**	1	0	1	1	1	1
phish	**BLR**	1	0	1	1	1	1
Wghtd	**BLR**	1	0	1	1	1	1
legit	**J48**	1	0.002	0.999	1	1	0.999
phish	**J48**	0.998	0	1	0.998	0.999	0.999
Wghtd	**J48**	0.999	0.001	0.999	0.999	0.999	0.999
legit	**NB**	0.999	0.074	0.968	0.999	0.983	0.971
phish	**NB**	0.926	0.001	0.999	0.926	0.961	0.979
Wghtd	**NB**	0.977	0.051	0.977	0.977	0.976	0.973
legit	**RF**	1	0.001	1	1	1	1
phish	**RF**	0.999	0	1	0.999	0.999	1
Wghtd	**RF**	1	0.001	1	1	1	1
legit	**SMO**	1	0	1	1	1	1
phish	**SMO**	1	0	1	1	1	1
Wghtd	**SMO**	1	0	1	1	1	1

Tables 5.89 to 5.98 summarize the results of the weka experiments on the combined dataset CSDMCRVL+NPNRV having full domains. Results till 10-grams are reported here. These results show that the detection rates are fairly constant for all n-grams.

Tables 5.99 to 5.108 summarize the results of the weka experiments on the combined dataset CSDMCRVL+NPNRV having domains with no TLD. Results till 10-grams are reported here. These results show that the detection rates are fairly constant for all n-grams.

Tables 5.109 and 5.110 summarize the results of the Confidence Weighted Algorithm experiments on the combined dataset CSDMCRVL+NPNRV having domains with no TLD. Results till 10-grams are reported here. These results show that the detection rates are fairly constant for all n-grams.

5.7 Information Gain

Since n-gram analysis results in creation of a very high number of features, we decided to find out which features were the most contributing ones towards the final classification. We selected the top 25 features according to their information gain values. Tables 5.111 and 5.112 show the features with their respective information gain values. These values clearly show a pattern that the domains with uh.edu in them were easily separable from the others. This is because of a large number of emails of the contributor having this particular domain segment.

Table 5.93: Results for 5-gram features of the full domains dataset

5gramFullDomainsFromHeader&IPsNoMsgIdCSDMCRVL+NPNRV							
Class	**Classifier**	**TPR**	**FPR**	**Precision**	**Recall**	**FMeasure**	**ROC**
legit	**ABoost**	0.999	0.004	0.998	0.999	0.999	1
phish	**ABoost**	0.996	0.001	0.998	0.996	0.997	1
Wghtd	**ABoost**	0.998	0.003	0.998	0.998	0.998	1
legit	**AttSel**	1	0.004	0.998	1	0.999	0.998
phish	**AttSel**	0.996	0	0.999	0.996	0.998	0.998
Wghtd	**AttSel**	0.999	0.003	0.999	0.999	0.999	0.998
legit	**Bagging**	1	0.001	1	1	1	1
phish	**Bagging**	0.999	0	1	0.999	0.999	1
Wghtd	**Bagging**	1	0.001	1	1	1	1
legit	**BLR**	1	0	1	1	1	1
phish	**BLR**	1	0	1	1	1	1
Wghtd	**BLR**	1	0	1	1	1	1
legit	**J48**	1	0.002	0.999	1	1	0.999
phish	**J48**	0.998	0	1	0.998	0.999	0.999
Wghtd	**J48**	0.999	0.001	0.999	0.999	0.999	0.999
legit	**NB**	0.999	0.074	0.968	0.999	0.983	0.971
phish	**NB**	0.926	0.001	0.999	0.926	0.961	0.979
Wghtd	**NB**	0.977	0.051	0.977	0.977	0.976	0.973
legit	**RF**	1	0.001	0.999	1	1	1
phish	**RF**	0.999	0	1	0.999	0.999	1
Wghtd	**RF**	1	0.001	1	1	1	1
legit	**SMO**	1	0	1	1	1	1
phish	**SMO**	1	0	1	1	1	1
Wghtd	**SMO**	1	0	1	1	1	1

Table 5.94: Results for 6-gram features of the full domains dataset

6gramFullDomainsFromHeader&IPsNoMsgIdCSDMCRVL+NPNRV							
Class	**Classifier**	**TPR**	**FPR**	**Precision**	**Recall**	**FMeasure**	**ROC**
legit	**ABoost**	0.999	0.004	0.998	0.999	0.999	1
phish	**ABoost**	0.996	0.001	0.998	0.996	0.997	1
Wghtd	**ABoost**	0.998	0.003	0.998	0.998	0.998	1
legit	**AttSel**	1	0.004	0.998	1	0.999	0.998
phish	**AttSel**	0.996	0	0.999	0.996	0.998	0.998
Wghtd	**AttSel**	0.999	0.003	0.999	0.999	0.999	0.998
legit	**Bagging**	1	0.001	1	1	1	1
phish	**Bagging**	0.999	0	1	0.999	0.999	1
Wghtd	**Bagging**	1	0.001	1	1	1	1
legit	**BLR**	1	0	1	1	1	1
phish	**BLR**	1	0	1	1	1	1
Wghtd	**BLR**	1	0	1	1	1	1
legit	**J48**	1	0.002	0.999	1	1	0.999
phish	**J48**	0.998	0	1	0.998	0.999	0.999
Wghtd	**J48**	0.999	0.001	0.999	0.999	0.999	0.999
legit	**NB**	1	0.074	0.968	1	0.983	0.971
phish	**NB**	0.926	0	0.999	0.926	0.961	0.979
Wghtd	**NB**	0.977	0.051	0.978	0.977	0.977	0.973
legit	**RF**	1	0.001	1	1	1	1
phish	**RF**	0.999	0	1	0.999	0.999	1
Wghtd	**RF**	1	0.001	1	1	1	1
legit	**SMO**	1	0	1	1	1	1
phish	**SMO**	1	0	1	1	1	1
Wghtd	**SMO**	1	0	1	1	1	1

Table 5.95: Results for 7-gram features of the full domains dataset

7gramFullDomainsFromHeader&IPsNoMsgIdCSDMCRVL+NPNRV							
Class	**Classifier**	**TPR**	**FPR**	**Precision**	**Recall**	**FMeasure**	**ROC**
legit	**ABoost**	0.999	0.004	0.998	0.999	0.999	1
phish	**ABoost**	0.996	0.001	0.998	0.996	0.997	1
Wghtd	**ABoost**	0.998	0.003	0.998	0.998	0.998	1
legit	**AttSel**	1	0.004	0.998	1	0.999	0.998
phish	**AttSel**	0.996	0	0.999	0.996	0.998	0.998
Wghtd	**AttSel**	0.999	0.003	0.999	0.999	0.999	0.998
legit	**Bagging**	1	0.001	1	1	1	1
phish	**Bagging**	0.999	0	1	0.999	0.999	1
Wghtd	**Bagging**	1	0.001	1	1	1	1
legit	**BLR**	1	0	1	1	1	1
phish	**BLR**	1	0	1	1	1	1
Wghtd	**BLR**	1	0	1	1	1	1
legit	**J48**	1	0.002	0.999	1	1	0.999
phish	**J48**	0.998	0	1	0.998	0.999	0.999
Wghtd	**J48**	0.999	0.001	0.999	0.999	0.999	0.999
legit	**NB**	1	0.074	0.968	1	0.983	0.971
phish	**NB**	0.926	0	0.999	0.926	0.961	0.979
Wghtd	**NB**	0.977	0.051	0.978	0.977	0.977	0.973
legit	**RF**	1	0	1	1	1	1
phish	**RF**	1	0	1	1	1	1
Wghtd	**RF**	1	0	1	1	1	1
legit	**SMO**	1	0	1	1	1	1
phish	**SMO**	1	0	1	1	1	1
Wghtd	**SMO**	1	0	1	1	1	1

Table 5.96: Results for 8-gram features of the full domains dataset

8gramFullDomainsFromHeader&IPsNoMsgIdCSDMCRVL+NPNRV							
Class	**Classifier**	**TPR**	**FPR**	**Precision**	**Recall**	**FMeasure**	**ROC**
legit	**ABoost**	0.999	0.004	0.998	0.999	0.999	1
phish	**ABoost**	0.996	0.001	0.998	0.996	0.997	1
Wghtd	**ABoost**	0.998	0.003	0.998	0.998	0.998	1
legit	**AttSel**	1	0.004	0.998	1	0.999	0.998
phish	**AttSel**	0.996	0	0.999	0.996	0.998	0.998
Wghtd	**AttSel**	0.999	0.003	0.999	0.999	0.999	0.998
legit	**Bagging**	1	0.001	1	1	1	1
phish	**Bagging**	0.999	0	1	0.999	0.999	1
Wghtd	**Bagging**	1	0.001	1	1	1	1
legit	**BLR**	1	0	1	1	1	1
phish	**BLR**	1	0	1	1	1	1
Wghtd	**BLR**	1	0	1	1	1	1
legit	**J48**	1	0.002	0.999	1	1	0.999
phish	**J48**	0.998	0	1	0.998	0.999	0.999
Wghtd	**J48**	0.999	0.001	0.999	0.999	0.999	0.999
legit	**NB**	1	0.074	0.968	1	0.983	0.971
phish	**NB**	0.926	0	0.999	0.926	0.961	0.979
Wghtd	**NB**	0.977	0.051	0.978	0.977	0.977	0.973
legit	**RF**	1	0.001	1	1	1	1
phish	**RF**	0.999	0	1	0.999	1	1
Wghtd	**RF**	1	0	1	1	1	1
legit	**SMO**	1	0	1	1	1	1
phish	**SMO**	1	0	1	1	1	1
Wghtd	**SMO**	1	0	1	1	1	1

Table 5.97: Results for 9-gram features of the full domains dataset

9gramFullDomainsFromHeader&IPsNoMsgIdCSDMCRVL+NPNRV							
Class	**Classifier**	**TPR**	**FPR**	**Precision**	**Recall**	**FMeasure**	**ROC**
legit	**ABoost**	0.999	0.004	0.998	0.999	0.999	1
phish	**ABoost**	0.996	0.001	0.998	0.996	0.997	1
Wghtd	**ABoost**	0.998	0.003	0.998	0.998	0.998	1
legit	**AttSel**	1	0.004	0.998	1	0.999	0.998
phish	**AttSel**	0.996	0	0.999	0.996	0.998	0.998
Wghtd	**AttSel**	0.999	0.003	0.999	0.999	0.999	0.998
legit	**Bagging**	1	0.001	1	1	1	1
phish	**Bagging**	0.999	0	1	0.999	0.999	1
Wghtd	**Bagging**	1	0.001	1	1	1	1
legit	**BLR**	1	0	1	1	1	1
phish	**BLR**	1	0	1	1	1	1
Wghtd	**BLR**	1	0	1	1	1	1
legit	**J48**	1	0.002	0.999	1	1	0.999
phish	**J48**	0.998	0	1	0.998	0.999	0.999
Wghtd	**J48**	0.999	0.001	0.999	0.999	0.999	0.999
legit	**NB**	1	0.074	0.968	1	0.983	0.971
phish	**NB**	0.926	0	0.999	0.926	0.961	0.979
Wghtd	**NB**	0.977	0.051	0.978	0.977	0.977	0.973
legit	**RF**	1	0.001	1	1	1	1
phish	**RF**	0.999	0	1	0.999	1	1
Wghtd	**RF**	1	0	1	1	1	1
legit	**SMO**	1	0	1	1	1	1
phish	**SMO**	1	0	1	1	1	1
Wghtd	**SMO**	1	0	1	1	1	1

Table 5.98: Results for 10-gram features of the full domains dataset

10gramFullDomainsFromHeader&IPsNoMsgIdCSDMCRVL+NPNRV							
Class	**Classifier**	**TPR**	**FPR**	**Precision**	**Recall**	**FMeasure**	**ROC**
legit	**ABoost**	0.999	0.004	0.998	0.999	0.999	1
phish	**ABoost**	0.996	0.001	0.998	0.996	0.997	1
Wghtd	**ABoost**	0.998	0.003	0.998	0.998	0.998	1
legit	**AttSel**	1	0.004	0.998	1	0.999	0.998
phish	**AttSel**	0.996	0	0.999	0.996	0.998	0.998
Wghtd	**AttSel**	0.999	0.003	0.999	0.999	0.999	0.998
legit	**Bagging**	1	0.001	1	1	1	1
phish	**Bagging**	0.999	0	1	0.999	0.999	1
Wghtd	**Bagging**	1	0.001	1	1	1	1
legit	**BLR**	1	0	1	1	1	1
phish	**BLR**	1	0	1	1	1	1
Wghtd	**BLR**	1	0	1	1	1	1
legit	**J48**	1	0.002	0.999	1	1	0.999
phish	**J48**	0.998	0	1	0.998	0.999	0.999
Wghtd	**J48**	0.999	0.001	0.999	0.999	0.999	0.999
legit	**NB**	1	0.074	0.968	1	0.983	0.971
phish	**NB**	0.926	0	0.999	0.926	0.961	0.979
Wghtd	**NB**	0.977	0.051	0.978	0.977	0.977	0.973
legit	**RF**	1	0	1	1	1	1
phish	**RF**	1	0	1	1	1	1
Wghtd	**RF**	1	0	1	1	1	1
legit	**SMO**	1	0	1	1	1	1
phish	**SMO**	1	0	1	1	1	1
Wghtd	**SMO**	1	0	1	1	1	1

Table 5.99: Results for 1-gram features of the domains with no TLD dataset

1gramNoTLDDomainsFromHeader&IPsNoMsgIdCSDMCRVL+NPNRV							
Class	**Classifier**	**TPR**	**FPR**	**Precision**	**Recall**	**FMeasure**	**ROC**
legit	**ABoost**	0.999	0.074	0.967	0.999	0.983	0.98
phish	**ABoost**	0.926	0.001	0.997	0.926	0.96	0.98
Wghtd	**ABoost**	0.976	0.052	0.977	0.976	0.976	0.98
legit	**AttSel**	0.997	0.059	0.974	0.997	0.985	0.969
phish	**AttSel**	0.941	0.003	0.992	0.941	0.966	0.969
Wghtd	**AttSel**	0.979	0.042	0.98	0.979	0.979	0.969
legit	**Bagging**	1	0.056	0.975	1	0.987	0.982
phish	**Bagging**	0.944	0	0.999	0.944	0.971	0.982
Wghtd	**Bagging**	0.982	0.039	0.983	0.982	0.982	0.982
legit	**BLR**	0.999	0.059	0.974	0.999	0.987	0.97
phish	**BLR**	0.941	0.001	0.998	0.941	0.969	0.97
Wghtd	**BLR**	0.981	0.041	0.982	0.981	0.981	0.97
legit	**J48**	0.999	0.056	0.975	0.999	0.987	0.981
phish	**J48**	0.944	0.001	0.997	0.944	0.97	0.981
Wghtd	**J48**	0.982	0.039	0.982	0.982	0.981	0.981
legit	**NB**	0.996	0.074	0.968	0.996	0.981	0.966
phish	**NB**	0.926	0.004	0.99	0.926	0.957	0.96
Wghtd	**NB**	0.974	0.052	0.975	0.974	0.974	0.964
legit	**RF**	1	0.055	0.976	1	0.988	0.983
phish	**RF**	0.945	0	1	0.945	0.971	0.983
Wghtd	**RF**	0.983	0.038	0.983	0.983	0.983	0.983
legit	**SMO**	0.999	0.055	0.976	0.999	0.987	0.972
phish	**SMO**	0.945	0.001	0.998	0.945	0.971	0.972
Wghtd	**SMO**	0.982	0.038	0.983	0.982	0.982	0.972

Table 5.100: Results for 2-gram features of the domains with no TLD dataset

2gramNoTLDDomainsFromHeader&IPsNoMsgIdCSDMCRVL+NPNRV							
Class	**Classifier**	**TPR**	**FPR**	**Precision**	**Recall**	**FMeasure**	**ROC**
legit	**ABoost**	0.997	0.059	0.974	0.997	0.985	0.98
phish	**ABoost**	0.941	0.003	0.993	0.941	0.966	0.98
Wghtd	**ABoost**	0.98	0.041	0.98	0.98	0.979	0.98
legit	**AttSel**	1	0.059	0.974	1	0.987	0.971
phish	**AttSel**	0.941	0	1	0.941	0.97	0.971
Wghtd	**AttSel**	0.982	0.04	0.982	0.982	0.982	0.971
legit	**Bagging**	1	0.056	0.975	1	0.987	0.983
phish	**Bagging**	0.944	0	0.999	0.944	0.971	0.983
Wghtd	**Bagging**	0.982	0.039	0.983	0.982	0.982	0.983
legit	**BLR**	1	0.055	0.976	1	0.988	0.973
phish	**BLR**	0.945	0	1	0.945	0.972	0.973
Wghtd	**BLR**	0.983	0.038	0.983	0.983	0.983	0.973
legit	**J48**	1	0.056	0.975	1	0.987	0.978
phish	**J48**	0.944	0	1	0.944	0.971	0.978
Wghtd	**J48**	0.983	0.039	0.983	0.983	0.982	0.978
legit	**NB**	0.995	0.074	0.968	0.995	0.981	0.97
phish	**NB**	0.926	0.005	0.989	0.926	0.956	0.962
Wghtd	**NB**	0.974	0.052	0.974	0.974	0.973	0.968
legit	**RF**	1	0.055	0.976	1	0.988	0.983
phish	**RF**	0.945	0	1	0.945	0.971	0.983
Wghtd	**RF**	0.983	0.038	0.983	0.983	0.983	0.983
legit	**SMO**	1	0.055	0.976	1	0.988	0.973
phish	**SMO**	0.945	0	1	0.945	0.972	0.973
Wghtd	**SMO**	0.983	0.038	0.983	0.983	0.983	0.973

Table 5.101: Results for 3-gram features of the domains with no TLD dataset

3gramNoTLDDomainsFromHeader&IPsNoMsgIdCSDMCRVL+NPNRV							
Class	**Classifier**	**TPR**	**FPR**	**Precision**	**Recall**	**FMeasure**	**ROC**
legit	**ABoost**	0.997	0.059	0.974	0.997	0.985	0.98
phish	**ABoost**	0.941	0.003	0.993	0.941	0.966	0.98
Wghtd	**ABoost**	0.98	0.041	0.98	0.98	0.979	0.98
legit	**AttSel**	1	0.059	0.974	1	0.987	0.971
phish	**AttSel**	0.941	0	1	0.941	0.97	0.971
Wghtd	**AttSel**	0.982	0.04	0.982	0.982	0.982	0.971
legit	**Bagging**	1	0.056	0.975	1	0.987	0.983
phish	**Bagging**	0.944	0	0.999	0.944	0.971	0.983
Wghtd	**Bagging**	0.982	0.039	0.983	0.982	0.982	0.983
legit	**BLR**	1	0.055	0.976	1	0.988	0.973
phish	**BLR**	0.945	0	1	0.945	0.972	0.973
Wghtd	**BLR**	0.983	0.038	0.983	0.983	0.983	0.973
legit	**J48**	1	0.056	0.975	1	0.987	0.978
phish	**J48**	0.944	0	1	0.944	0.971	0.978
Wghtd	**J48**	0.983	0.039	0.983	0.983	0.982	0.978
legit	**NB**	0.995	0.074	0.968	0.995	0.981	0.971
phish	**NB**	0.926	0.005	0.989	0.926	0.956	0.962
Wghtd	**NB**	0.974	0.052	0.974	0.974	0.973	0.968
legit	**RF**	1	0.055	0.976	1	0.988	0.983
phish	**RF**	0.945	0	1	0.945	0.971	0.983
Wghtd	**RF**	0.983	0.038	0.983	0.983	0.983	0.983
legit	**SMO**	1	0.055	0.976	1	0.988	0.973
phish	**SMO**	0.945	0	1	0.945	0.972	0.973
Wghtd	**SMO**	0.983	0.038	0.983	0.983	0.983	0.973

Table 5.102: Results for 4-gram features of the domains with no TLD dataset

4gramNoTLDDomainsFromHeader&IPsNoMsgIdCSDMCRVL+NPNRV							
Class	**Classifier**	**TPR**	**FPR**	**Precision**	**Recall**	**FMeasure**	**ROC**
legit	**ABoost**	1	0.055	0.976	1	0.988	0.973
phish	**ABoost**	0.945	0	1	0.945	0.972	0.973
Wghtd	**ABoost**	0.983	0.038	0.983	0.983	0.983	0.973
legit	**AttSel**	1	0.059	0.974	1	0.987	0.971
phish	**AttSel**	0.941	0	1	0.941	0.97	0.971
Wghtd	**AttSel**	0.982	0.04	0.982	0.982	0.982	0.971
legit	**Bagging**	1	0.056	0.975	1	0.987	0.983
phish	**Bagging**	0.944	0	0.999	0.944	0.971	0.983
Wghtd	**Bagging**	0.982	0.039	0.983	0.982	0.982	0.983
legit	**BLR**	1	0.055	0.976	1	0.988	0.973
phish	**BLR**	0.945	0	1	0.945	0.972	0.973
Wghtd	**BLR**	0.983	0.038	0.983	0.983	0.983	0.973
legit	**J48**	1	0.056	0.975	1	0.987	0.978
phish	**J48**	0.944	0	1	0.944	0.971	0.978
Wghtd	**J48**	0.983	0.039	0.983	0.983	0.982	0.978
legit	**NB**	0.995	0.074	0.968	0.995	0.981	0.97
phish	**NB**	0.926	0.005	0.989	0.926	0.956	0.961
Wghtd	**NB**	0.974	0.052	0.974	0.974	0.973	0.967
legit	**RF**	1	0.055	0.976	1	0.988	0.983
phish	**RF**	0.945	0	1	0.945	0.971	0.983
Wghtd	**RF**	0.983	0.038	0.983	0.983	0.983	0.983
legit	**SMO**	1	0.055	0.976	1	0.988	0.973
phish	**SMO**	0.945	0	1	0.945	0.972	0.973
Wghtd	**SMO**	0.983	0.038	0.983	0.983	0.983	0.973

Table 5.103: Results for 5-gram features of the domains with no TLD dataset

5gramNoTLDDomainsFromHeader&IPsNoMsgIdCSDMCRVL+NPNRV							
Class	**Classifier**	**TPR**	**FPR**	**Precision**	**Recall**	**FMeasure**	**ROC**
legit	**ABoost**	0.997	0.059	0.974	0.997	0.985	0.98
phish	**ABoost**	0.941	0.003	0.993	0.941	0.966	0.98
Wghtd	**ABoost**	0.98	0.041	0.98	0.98	0.979	0.98
legit	**AttSel**	1	0.059	0.974	1	0.987	0.971
phish	**AttSel**	0.941	0	1	0.941	0.97	0.971
Wghtd	**AttSel**	0.982	0.04	0.982	0.982	0.982	0.971
legit	**Bagging**	1	0.056	0.975	1	0.987	0.983
phish	**Bagging**	0.944	0	0.999	0.944	0.971	0.983
Wghtd	**Bagging**	0.982	0.039	0.983	0.982	0.982	0.983
legit	**BLR**	1	0.055	0.976	1	0.988	0.973
phish	**BLR**	0.945	0	1	0.945	0.972	0.973
Wghtd	**BLR**	0.983	0.038	0.983	0.983	0.983	0.973
legit	**J48**	1	0.056	0.975	1	0.987	0.978
phish	**J48**	0.944	0	1	0.944	0.971	0.978
Wghtd	**J48**	0.983	0.039	0.983	0.983	0.982	0.978
legit	**NB**	0.995	0.074	0.968	0.995	0.981	0.97
phish	**NB**	0.926	0.005	0.989	0.926	0.956	0.961
Wghtd	**NB**	0.974	0.052	0.974	0.974	0.973	0.967
legit	**RF**	1	0.055	0.976	1	0.988	0.983
phish	**RF**	0.945	0	1	0.945	0.972	0.983
Wghtd	**RF**	0.983	0.038	0.983	0.983	0.983	0.983
legit	**SMO**	1	0.055	0.976	1	0.988	0.973
phish	**SMO**	0.945	0	1	0.945	0.972	0.973
Wghtd	**SMO**	0.983	0.038	0.983	0.983	0.983	0.973

Table 5.104: Results for 6-gram features of the domains with no TLD dataset

6gramNoTLDDomainsFromHeader&IPsNoMsgIdCSDMCRVL+NPNRV							
Class	**Classifier**	**TPR**	**FPR**	**Precision**	**Recall**	**FMeasure**	**ROC**
legit	**ABoost**	0.997	0.059	0.974	0.997	0.985	0.98
phish	**ABoost**	0.941	0.003	0.993	0.941	0.966	0.98
Wghtd	**ABoost**	0.98	0.041	0.98	0.98	0.979	0.98
legit	**AttSel**	1	0.059	0.974	1	0.987	0.971
phish	**AttSel**	0.941	0	1	0.941	0.97	0.971
Wghtd	**AttSel**	0.982	0.04	0.982	0.982	0.982	0.971
legit	**Bagging**	1	0.056	0.975	1	0.987	0.983
phish	**Bagging**	0.944	0	0.999	0.944	0.971	0.983
Wghtd	**Bagging**	0.982	0.039	0.983	0.982	0.982	0.983
legit	**BLR**	1	0.055	0.976	1	0.988	0.973
phish	**BLR**	0.945	0	1	0.945	0.972	0.973
Wghtd	**BLR**	0.983	0.038	0.983	0.983	0.983	0.973
legit	**J48**	1	0.056	0.975	1	0.987	0.978
phish	**J48**	0.944	0	1	0.944	0.971	0.978
Wghtd	**J48**	0.983	0.039	0.983	0.983	0.982	0.978
legit	**NB**	0.995	0.074	0.968	0.995	0.981	0.97
phish	**NB**	0.926	0.005	0.989	0.926	0.956	0.961
Wghtd	**NB**	0.974	0.052	0.974	0.974	0.973	0.967
legit	**RF**	1	0.055	0.976	1	0.988	0.983
phish	**RF**	0.945	0	1	0.945	0.972	0.983
Wghtd	**RF**	0.983	0.038	0.983	0.983	0.983	0.983
legit	**SMO**	1	0.055	0.976	1	0.988	0.973
phish	**SMO**	0.945	0	1	0.945	0.972	0.973
Wghtd	**SMO**	0.983	0.038	0.983	0.983	0.983	0.973

Table 5.105: Results for 7-gram features of the domains with no TLD dataset

7gramNoTLDDomainsFromHeader&IPsNoMsgIdCSDMCRVL+NPNRV							
Class	**Classifier**	**TPR**	**FPR**	**Precision**	**Recall**	**FMeasure**	**ROC**
legit	**ABoost**	0.997	0.059	0.974	0.997	0.985	0.98
phish	**ABoost**	0.941	0.003	0.993	0.941	0.966	0.98
Wghtd	**ABoost**	0.98	0.041	0.98	0.98	0.979	0.98
legit	**AttSel**	1	0.059	0.974	1	0.987	0.971
phish	**AttSel**	0.941	0	1	0.941	0.97	0.971
Wghtd	**AttSel**	0.982	0.04	0.982	0.982	0.982	0.971
legit	**Bagging**	1	0.056	0.975	1	0.987	0.983
phish	**Bagging**	0.944	0	0.999	0.944	0.971	0.983
Wghtd	**Bagging**	0.982	0.039	0.983	0.982	0.982	0.983
legit	**BLR**	1	0.055	0.976	1	0.988	0.973
phish	**BLR**	0.945	0	1	0.945	0.972	0.973
Wghtd	**BLR**	0.983	0.038	0.983	0.983	0.983	0.973
legit	**J48**	1	0.056	0.975	1	0.987	0.978
phish	**J48**	0.944	0	1	0.944	0.971	0.978
Wghtd	**J48**	0.983	0.039	0.983	0.983	0.982	0.978
legit	**NB**	0.995	0.074	0.968	0.995	0.981	0.97
phish	**NB**	0.926	0.005	0.989	0.926	0.956	0.961
Wghtd	**NB**	0.974	0.052	0.974	0.974	0.973	0.967
legit	**RF**	1	0.055	0.976	1	0.988	0.983
phish	**RF**	0.945	0	1	0.945	0.972	0.983
Wghtd	**RF**	0.983	0.038	0.983	0.983	0.983	0.983
legit	**SMO**	1	0.055	0.976	1	0.988	0.973
phish	**SMO**	0.945	0	1	0.945	0.972	0.973
Wghtd	**SMO**	0.983	0.038	0.983	0.983	0.983	0.973

Table 5.106: Results for 8-gram features of the domains with no TLD dataset

8gramNoTLDDomainsFromHeader&IPsNoMsgIdCSDMCRVL+NPNRV							
Class	**Classifier**	**TPR**	**FPR**	**Precision**	**Recall**	**FMeasure**	**ROC**
legit	**ABoost**	0.997	0.059	0.974	0.997	0.985	0.98
phish	**ABoost**	0.941	0.003	0.993	0.941	0.966	0.98
Wghtd	**ABoost**	0.98	0.041	0.98	0.98	0.979	0.98
legit	**AttSel**	1	0.059	0.974	1	0.987	0.971
phish	**AttSel**	0.941	0	1	0.941	0.97	0.971
Wghtd	**AttSel**	0.982	0.04	0.982	0.982	0.982	0.971
legit	**BLR**	1	0.055	0.976	1	0.988	0.973
phish	**BLR**	0.945	0	1	0.945	0.972	0.973
Wghtd	**BLR**	0.983	0.038	0.983	0.983	0.983	0.973
legit	**J48**	1	0.056	0.975	1	0.987	0.978
phish	**J48**	0.944	0	1	0.944	0.971	0.978
Wghtd	**J48**	0.983	0.039	0.983	0.983	0.982	0.978
legit	**NB**	0.995	0.074	0.968	0.995	0.981	0.97
phish	**NB**	0.926	0.005	0.989	0.926	0.956	0.961
Wghtd	**NB**	0.974	0.052	0.974	0.974	0.973	0.967
legit	**RF**	1	0.055	0.976	1	0.988	0.983
phish	**RF**	0.945	0	1	0.945	0.972	0.983
Wghtd	**RF**	0.983	0.038	0.983	0.983	0.983	0.983
legit	**SMO**	1	0.055	0.976	1	0.988	0.973
phish	**SMO**	0.945	0	1	0.945	0.972	0.973
Wghtd	**SMO**	0.983	0.038	0.983	0.983	0.983	0.973

Table 5.107: Results for 9-gram features of the domains with no TLD dataset

9gramNoTLDDomainsFromHeader&IPsNoMsgIdCSDMCRVL+NPNRV							
Class	**Classifier**	**TPR**	**FPR**	**Precision**	**Recall**	**FMeasure**	**ROC**
legit	**ABoost**	0.997	0.059	0.974	0.997	0.985	0.98
phish	**ABoost**	0.941	0.003	0.993	0.941	0.966	0.98
Wghtd	**ABoost**	0.98	0.041	0.98	0.98	0.979	0.98
legit	**AttSel**	1	0.059	0.974	1	0.987	0.971
phish	**AttSel**	0.941	0	1	0.941	0.97	0.971
Wghtd	**AttSel**	0.982	0.04	0.982	0.982	0.982	0.971
legit	**BLR**	1	0.055	0.976	1	0.988	0.973
phish	**BLR**	0.945	0	1	0.945	0.972	0.973
Wghtd	**BLR**	0.983	0.038	0.983	0.983	0.983	0.973
legit	**J48**	1	0.056	0.975	1	0.987	0.978
phish	**J48**	0.944	0	1	0.944	0.971	0.978
Wghtd	**J48**	0.983	0.039	0.983	0.983	0.982	0.978
legit	**NB**	0.995	0.074	0.968	0.995	0.981	0.97
phish	**NB**	0.926	0.005	0.989	0.926	0.956	0.961
Wghtd	**NB**	0.974	0.052	0.974	0.974	0.973	0.967
legit	**RF**	1	0.055	0.976	1	0.988	0.983
phish	**RF**	0.945	0	1	0.945	0.972	0.983
Wghtd	**RF**	0.983	0.038	0.983	0.983	0.983	0.983
legit	**SMO**	1	0.055	0.976	1	0.988	0.973
phish	**SMO**	0.945	0	1	0.945	0.972	0.973
Wghtd	**SMO**	0.983	0.038	0.983	0.983	0.983	0.973

Table 5.108: Results for 10-gram features of the domains with no TLD dataset

10gramNoTLDDomainsFromHeader&IPsNoMsgIdCSDMCRVL+NPNRV							
Class	**Classifier**	**TPR**	**FPR**	**Precision**	**Recall**	**FMeasure**	**ROC**
legit	**ABoost**	0.997	0.059	0.974	0.997	0.985	0.98
phish	**ABoost**	0.941	0.003	0.993	0.941	0.966	0.98
Wghtd	**ABoost**	0.98	0.041	0.98	0.98	0.979	0.98
legit	**AttSel**	1	0.059	0.974	1	0.987	0.971
phish	**AttSel**	0.941	0	1	0.941	0.97	0.971
Wghtd	**AttSel**	0.982	0.04	0.982	0.982	0.982	0.971
legit	**BLR**	1	0.055	0.976	1	0.988	0.973
phish	**BLR**	0.945	0	1	0.945	0.972	0.973
Wghtd	**BLR**	0.983	0.038	0.983	0.983	0.983	0.973
legit	**J48**	1	0.056	0.975	1	0.987	0.978
phish	**J48**	0.944	0	1	0.944	0.971	0.978
Wghtd	**J48**	0.983	0.039	0.983	0.983	0.982	0.978
legit	**NB**	0.995	0.059	0.974	0.995	0.984	0.97
phish	**NB**	0.941	0.005	0.989	0.941	0.964	0.961
Wghtd	**NB**	0.978	0.042	0.979	0.978	0.978	0.967
legit	**SMO**	1	0.055	0.976	1	0.988	0.973
phish	**SMO**	0.945	0	1	0.945	0.972	0.973
Wghtd	**SMO**	0.983	0.038	0.983	0.983	0.983	0.973

Table 5.109: Confidence Weighted Algo Results for Full Domains

FullDomainsCSDMCRVL+NPNRV		
Gram	**TP Rate**	**FP Rate**
1	99.937	0.111
2	100	0.195
3	99.986	0.195
4	99.987	0.195
5	99.987	0.195
6	99.987	0.195
7	99.987	0.195
8	99.987	0.195
9	99.987	0.195
10	99.987	0.195

Table 5.110: Confidence Weighted Algo Results for Full Domains

NoTLDDomainsCSDMCRVL+NPNRV		
Gram	**TP Rate**	**FP Rate**
1	99.912	5.656
2	93.535	5.155
3	99.987	5.656
4	93.8	5.322
5	99.987	5.656
6	93.699	4.96
7	99.987	5.656
8	93.938	5.183
9	99.987	5.656
10	93.585	5.183

5.8 Comparative Analysis

We make a direct comparison with Phish-IDetector, a system which uses ngram analysis on Message-IDs of the emails. For this purpose we have extracted header domains excluding the Message-ID header for the header domain analysis and only the Message-IDs separately. Running experiments on the Message-IDs and the rest of the header domains reveal that our header domain analysis produces better detection and greatly reduces the false positive rate. We can thus, infer from these experiments that the header domains are a better indicator of legitimacy of emails than just the Message-ID.

There has been some work done in using the SMTP path of an email for classification [26] but it has only been used for spam detection and not for phishing. Besides, they only make use of the SMTP path as indicated by the IP addresses in the Received fields, whereas, we have collected the domains from several headers besides those like From, Delivered-To, CC etc.

Tables 5.113 and 5.116 summarize the results of the weka experiments on the combined dataset CSDMCRVL+NPNRV for both RHS and Split Message-ID n-gram features. Results till 2-grams are reported here. These results show that the detection rate and false positive rate are slightly better for Split Message-ID than that for RHS Message-ID datasets.

5.9 Error Analysis

It is very important to find out how any system can be improved further. We took a closer look at the misclassifications performed by our classifiers to find out the shortcomings our technique and what could be done to make it better.

We checked the misclassified emails for both full domains as well as domains with no TLDs and made some observations. A legitimate email was constantly marked phish as it did not have enough information in the header. The only domain available was "cs.uh.edu" and it was insufficient for proper classification.

For the phishing emails some of the wrongly classified emails did not have complete headers and hence ended up providing only single domain, for example, "paypal.com", "westernunion.com".

Another important observation was that removing TLDs from the domains caused information loss and increased the false positive rate. But the false negative rate was not affected significantly. This means that though some legitimate emails ended up being classified as phish, the phishing emails were classified as legitimate, which is of more importance for phishing classifiers. The cost associate with false negatives is much higher than that associated with false positives.

5.10 Security Analysis

As mentioned, the exponentially increasing file size for higher order n-grams makes it difficult to run different classifiers on them without using specialized big data

approaches. We currently ran only the confidence weighted algorithm on higher order n-gram files, which has proven itself to be competitive in other scenarios, but not guaranteed to be the ideal choice for best results. Phishers could try to obfuscate the header domains and try to evade our system, however they cannot change the entire path of the email. For instance, the Received-From headers closest to the receiver's end are not under the control of the sender. Also, the combination of the domains of the Received-From headers and the other headers would help in case of such obfuscations.

Aggregating the header domain analysis with the SMTP features would also help in identifying such cases of obfuscation.

Table 5.111: Information Gain Values for 5gramFullDomainsBalNazarioPhish-NewRVL

5gramFullDomainsBalNazarioPhishNewRVL		
S.No.	**IG**	**Feature**
1	0.985588	o
2	0.98479	m
3	0.98304	space
4	0.98304	“space
5	0.9811	or
6	0.975533	u
7	0.973901	edu
8	0.973901	.ed
9	0.973901	du
10	0.973901	.e
11	0.973901	.edu
12	0.969608	h.
13	0.969608	uh
14	0.969608	h.e
15	0.969608	uh.e
16	0.969608	uh.ed
17	0.969608	h.ed
18	0.969608	h.edu
19	0.969608	uh.
20	0.968558	s.
21	0.967529	ail
22	0.967529	mai
23	0.967529	mail
24	0.967529	il
25	0.967529	ai

Table 5.112: Information Gain Values for 5gramNoTLDDomainsBalNazarioPhish-NewRVL

5gramNoTLDDomainsBalNazarioPhishNewRVL		
S.No.	**IG**	**Feature**
1	0.977489	e
2	0.977009	m
3	0.973901	uh
4	0.973352	c
5	0.972813	cs.
6	0.972813	cs
7	0.972813	.u
8	0.972813	.uh
9	0.972813	s.
10	0.972266	o
11	0.969608	cs.uh
12	0.969608	s.u
13	0.969608	s.uh
14	0.969608	cs.u
15	0.96831	s
16	0.967529	il
17	0.967529	ail
18	0.967529	ai
19	0.967529	mail
20	0.967529	mai
21	0.966257	ma
22	0.965111	u
23	0.963732	l
24	0.963428	.cs.
25	0.963428	.cs

Table 5.113: Results for 1-gram features of the RHS Message-IDs from combined dataset

1gramRHSCSDMCRVL+NPNRV							
Class	**Classifier**	**TPR**	**FPR**	**Precision**	**Recall**	**FMeasure**	**ROC**
legit	**ABoost**	0.94	0.725	0.742	0.94	0.829	0.727
phish	**ABoost**	0.275	0.06	0.676	0.275	0.391	0.727
Wghtd	**ABoost**	0.733	0.518	0.721	0.733	0.693	0.727
legit	**AttSel**	0.959	0.351	0.858	0.959	0.906	0.905
phish	**AttSel**	0.649	0.041	0.878	0.649	0.746	0.905
Wghtd	**AttSel**	0.863	0.254	0.864	0.863	0.856	0.905
legit	**Bagging**	0.974	0.093	0.959	0.974	0.967	0.988
phish	**Bagging**	0.907	0.026	0.941	0.907	0.924	0.988
Wghtd	**Bagging**	0.954	0.072	0.953	0.954	0.953	0.988
legit	**BLR**	0.894	0.53	0.789	0.894	0.838	0.682
phish	**BLR**	0.47	0.106	0.668	0.47	0.552	0.682
Wghtd	**BLR**	0.762	0.398	0.751	0.762	0.749	0.682
legit	**J48**	0.976	0.071	0.968	0.976	0.972	0.984
phish	**J48**	0.929	0.024	0.946	0.929	0.938	0.984
Wghtd	**J48**	0.962	0.056	0.961	0.962	0.961	0.984
legit	**NB**	0.741	0.363	0.819	0.741	0.778	0.767
phish	**NB**	0.637	0.259	0.526	0.637	0.576	0.767
Wghtd	**NB**	0.709	0.33	0.728	0.709	0.715	0.767
legit	**RF**	0.999	0.026	0.988	0.999	0.993	0.998
phish	**RF**	0.974	0.001	0.997	0.974	0.985	0.998
Wghtd	**RF**	0.991	0.019	0.991	0.991	0.991	0.998
legit	**SMO**	0.914	0.583	0.776	0.914	0.84	0.666
phish	**SMO**	0.417	0.086	0.687	0.417	0.519	0.666
Wghtd	**SMO**	0.76	0.428	0.749	0.76	0.74	0.666

Table 5.114: Results for 2-gram features of the RHS Message-IDs from combined dataset

2gramRHSCSDMCRVL+NPNRV							
Class	**Classifier**	**TPR**	**FPR**	**Precision**	**Recall**	**FMeasure**	**ROC**
legit	**ABoost**	0.995	0.841	0.724	0.995	0.838	0.733
phish	**ABoost**	0.159	0.005	0.934	0.159	0.271	0.733
Wghtd	**ABoost**	0.735	0.581	0.789	0.735	0.662	0.733
legit	**AttSel**	0.94	0.468	0.816	0.94	0.874	0.848
phish	**AttSel**	0.532	0.06	0.799	0.532	0.639	0.848
Wghtd	**AttSel**	0.813	0.341	0.811	0.813	0.8	0.848
legit	**Bagging**	0.98	0.068	0.97	0.98	0.975	0.99
phish	**Bagging**	0.932	0.02	0.955	0.932	0.943	0.99
Wghtd	**Bagging**	0.965	0.053	0.965	0.965	0.965	0.99
legit	**BLR**	0.975	0.114	0.95	0.975	0.962	0.93
phish	**BLR**	0.886	0.025	0.94	0.886	0.913	0.93
Wghtd	**BLR**	0.947	0.086	0.947	0.947	0.947	0.93
legit	**J48**	0.984	0.053	0.976	0.984	0.98	0.992
phish	**J48**	0.947	0.016	0.964	0.947	0.956	0.992
Wghtd	**J48**	0.973	0.041	0.973	0.973	0.973	0.992
legit	**NB**	0.597	0.128	0.912	0.597	0.722	0.82
phish	**NB**	0.872	0.403	0.494	0.872	0.631	0.821
Wghtd	**NB**	0.683	0.213	0.782	0.683	0.693	0.82
legit	**RF**	0.999	0.026	0.988	0.999	0.994	0.998
phish	**RF**	0.974	0.001	0.997	0.974	0.985	0.998
Wghtd	**RF**	0.991	0.018	0.991	0.991	0.991	0.998
legit	**SMO**	0.973	0.137	0.94	0.973	0.956	0.918
phish	**SMO**	0.863	0.027	0.934	0.863	0.898	0.918
Wghtd	**SMO**	0.939	0.103	0.939	0.939	0.938	0.918

Table 5.115: Results for 1-gram features of the Split Message-IDs from combined dataset

1gramSplitMsgIdCSDMCRVL+NPNRV							
Class	**Classifier**	**TPR**	**FPR**	**Precision**	**Recall**	**FMeasure**	**ROC**
legit	**ABoost**	0.995	0.95	0.699	0.995	0.821	0.688
phish	**ABoost**	0.05	0.005	0.813	0.05	0.094	0.688
Wghtd	**ABoost**	0.701	0.656	0.734	0.701	0.595	0.688
legit	**AttSel**	0.934	0.589	0.778	0.934	0.849	0.81
phish	**AttSel**	0.411	0.066	0.737	0.411	0.528	0.81
Wghtd	**AttSel**	0.771	0.426	0.765	0.771	0.749	0.81
legit	**Bagging**	0.962	0.139	0.939	0.962	0.95	0.979
phish	**Bagging**	0.861	0.038	0.911	0.861	0.885	0.979
Wghtd	**Bagging**	0.93	0.108	0.93	0.93	0.93	0.979
legit	**BLR**	0.925	0.617	0.768	0.925	0.839	0.654
phish	**BLR**	0.383	0.075	0.697	0.383	0.494	0.654
Wghtd	**BLR**	0.756	0.449	0.746	0.756	0.732	0.654
legit	**J48**	0.965	0.103	0.954	0.965	0.959	0.97
phish	**J48**	0.897	0.035	0.92	0.897	0.908	0.97
Wghtd	**J48**	0.944	0.082	0.943	0.944	0.944	0.97
legit	**NB**	0.404	0.22	0.803	0.404	0.537	0.663
phish	**NB**	0.78	0.596	0.371	0.78	0.503	0.663
Wghtd	**NB**	0.521	0.337	0.668	0.521	0.527	0.663
legit	**RF**	0.997	0.02	0.991	0.997	0.994	1
phish	**RF**	0.98	0.003	0.994	0.98	0.987	1
Wghtd	**RF**	0.992	0.015	0.992	0.992	0.992	1
legit	**SMO**	0.981	0.825	0.725	0.981	0.834	0.578
phish	**SMO**	0.175	0.019	0.808	0.175	0.288	0.578
Wghtd	**SMO**	0.73	0.574	0.751	0.73	0.664	0.578

Table 5.116: Results for 2-gram features of the Split Message-IDs from combined dataset

2gramSplitMsgIdCSDMCRVL+NPNRV							
Class	**Classifier**	**TPR**	**FPR**	**Precision**	**Recall**	**FMeasure**	**ROC**
legit	**ABoost**	1	0.941	0.702	1	0.825	0.706
phish	**ABoost**	0.059	0	0.986	0.059	0.112	0.706
Wghtd	**ABoost**	0.707	0.648	0.79	0.707	0.603	0.706
legit	**AttSel**	0.931	0.358	0.852	0.931	0.89	0.894
phish	**AttSel**	0.642	0.069	0.808	0.642	0.716	0.894
Wghtd	**AttSel**	0.841	0.268	0.838	0.841	0.836	0.894
legit	**Bagging**	0.976	0.105	0.954	0.976	0.965	0.988
phish	**Bagging**	0.895	0.024	0.944	0.895	0.919	0.988
Wghtd	**Bagging**	0.951	0.08	0.951	0.951	0.951	0.988
legit	**BLR**	0.946	0.162	0.928	0.946	0.937	0.892
phish	**BLR**	0.838	0.054	0.875	0.838	0.856	0.892
Wghtd	**BLR**	0.912	0.128	0.912	0.912	0.912	0.892
legit	**J48**	0.981	0.058	0.974	0.981	0.977	0.99
phish	**J48**	0.942	0.019	0.957	0.942	0.95	0.99
Wghtd	**J48**	0.969	0.046	0.969	0.969	0.969	0.99
legit	**NB**	0.516	0.191	0.857	0.516	0.644	0.749
phish	**NB**	0.809	0.484	0.43	0.809	0.562	0.749
Wghtd	**NB**	0.607	0.282	0.724	0.607	0.619	0.749
legit	**RF**	0.999	0.022	0.99	0.999	0.994	1
phish	**RF**	0.978	0.001	0.997	0.978	0.987	1
Wghtd	**RF**	0.992	0.016	0.992	0.992	0.992	1
legit	**SMO**	0.95	0.166	0.927	0.95	0.938	0.892
phish	**SMO**	0.834	0.05	0.883	0.834	0.858	0.892
Wghtd	**SMO**	0.914	0.13	0.913	0.914	0.913	0.892

Chapter 6

SMTP Analysis

In this thesis, we attempt to go beyond the problem of email classification. The emails provide us with much more information and we extract these to find out about the SMTP servers involved in the email relaying process. We attempt to take a look at the state of the internet through the perspective of emails. We collect the statistics like the total number of SMTP servers for each email, which is a representation of the length of its path and the statistics about the percentage of SMTP servers open. The main aim of this experiment is to determine if the SMTP servers of domains in the phishing emails are more likely to be open than those of the legitimate emails. This information would give substantial proof to either support or reject our hypothesis of phishers using source routing. For every email all the Received-From domains are extracted and using nslookup command all the smtp servers are collected for each of those domains. Then the state of each of the smtp servers is checked using nmap command. Examples of the commands used are as follows:

nslookup -querytype=mx domainName

where domainName is the Received-From domain

nmap -p25 -PN smtpServer

where smtpServer is the smtp server returned by nslookup

Since this involves determining the state of the SMTP port on the server, and we use nmap command to do so, here is a list to get familiar with the port states returned by this command [27].

6.1 States Returned by nmap for SMTP Server

According to [27], the different states returned by nmap command are:

1. **open**

   This indicates that an application is actively accepting TCP connections, UDP datagrams or SCTP connections on this port. Usually, the purpose of port scanning is to determine which ports are open. Open ports are exploitable for security attacks. There is a constant conflict between the attackers an the administrators as the former tries to exploit and the latter tries to protect the ports. From a non-security point of view, these scans provide information

about the services that are present in the network for use.

2. **closed**

   This indicates that the port is accessible (it receives and responds to nmap probe packets), but no application is listening on it. This gives some information like a host is up on an IP address (host discovery, or ping scanning), and as part of OS detection. Though not active, closed ports are still reachable and might open in the future, which could be determined using port scanning. To block such ports from revealing any information, the administrators may use firewalls.

3. **filtered**

   This indicates that nmap cannot determine whether the port is open because its probes cannot reach the port because packet filtering prevents it. This could be due to a dedicated firewall device, router rules, or host-based firewall software. These ports are very frustrating for the attackers as they provide so little information because the filters simply drop probes without responding. Very rarely they respond with ICMP error messages such as type 3 code 13 (destination unreachable: communication administratively prohibited). Usually the probes are dropped, which slows down the scan dramatically. It forces nmap to retry several times to check if the probe was dropped due to network congestion rather than filtering.

4. **unfiltered**

   This indicates that a port is accessible, but nmap cannot determine whether it

is open or closed. Only the ACK scan, used to map firewall rulesets, designates ports into this state. Other scan types such as Window scan, SYN scan, or FIN scan, may provide information whether the port is open.

5. **open—filtered**

   This indicates that nmap is unable to determine whether a port is open or filtered. This happens for the scans where open ports give no response. The packet filter could have dropped the probe or any corresponding response so nmap does not know for sure whether the port is open or being filtered.

6. **closed—filtered**

   This indicates that nmap is unable to determine whether a port is closed or filtered. It is only used for the IP ID idle scan.

6.2 Three Options for SMTP State

The possible number of states being six in total, we had to decide on the aggregation of states of each of the SMTP servers for a domain. This would require converting the states to binary values of open (1) or close (0). We finally used three different options for the feature creation.

1. **Option 1**

   We keep these granular state information for each SMTP server returned for the Received-From domains in the header. We combine the name of the server

with its corresponding state and get n-grams from this combination and use these n-grams as the features.

2. **Option 2: The Strict-Open Assumption**

   Here, we perform the aggregation of the states of all the SMTP servers returned by each of the Received-From domains. We converted the states to binary form by assigning the state 1 to only those SMTP servers which returned 'open' state; all other 5 states are assigned 0. If any of the SMTP server had open state - 1, the corresponding domain was considered to have open state - 1. Again the domains were combined with the aggregated states and n-gram features were derived from them.

3. **Option 3: The Strict-Closed Assumption**

   Same as Option 2 but here the states were converted to binary form by assigning the state 0 to only those SMTP servers which returned 'close' state; all other 5 states are assigned 1.

6.3 Inference

Our experiments revealed that none of the SMTP servers were in "open" state. They were all either "filtered" or nmap failed to resolve the server host name. This shows that the system administrators are taking care not to leave any SMTP servers as open relays. However, [6] talks about more sophisticated ways of exploiting an SMTP server such that it acts as an open relay even when it is actually closed to outside traffic. A more specialized technique will have to be developed if we want to

find out about such exploitable SMTP servers.

6.4 SMTP Domains Intersection

The SMTP domains from different datasets had some intersections. For the pairs of datasets CSDMC+RV, CSDMC+RVL and RVL+NPN there were no intersecting domains. But for the pairs CSDMC+NPN, RVL+RV and NPN+RV the intersections have been listed in tables 6.1 - 6.3. The tables show that the frequencies of the intersecting domains between the legitimate (CSDMC) and phishing (NPN) sets are much lesser than the frequencies of the intersecting domains between the phishing sets NPN and RV. The numbers are high for the legitimate + phishing set of RVL+RV and this is because they are all from the same individual's inbox.

Table 6.1: Intersecting SMTP Domains for CSDMC+NPN

CSDMC+NPN	
SMTPServer	**TotalFrequency**
hotmail.com	28
eastrmimpo03.cox.net	19
eastrmimpo01.cox.net	18
eastrmimpo02.cox.net	16
eastrmmtao103.cox.net	12
eastrmmtao102.cox.net	11
edge03.upcmail.net	5
smtp.newsguy.com	4
free.fr	3
hrndva-omtalb.mail.rr.com	3
pih-relay04.plus.net	3
smtp-out4.blueyonder.co.uk	3
defout.telus.net	2
edge01.upcmail.net	2
fed1rmimpo03.cox.net	2
fed1rmmtao102.cox.net	2
filter.sfr.fr	2
mail02.svc.cra.dublin.eircom.net	2

Table 6.2: Intersecting SMTP Domains for RVL+RV

RVL+RV	
SMTPServer	**TotalFrequency**
dijkstra.cs.uh.edu	11989
smtp3.cc.uh.edu	2161
smtp4.cc.uh.edu	2089
localhost.localdomain	123
yahoogroups.com	24
tx2outboundpool.messaging.microsoft.com	5
edge01.upcmail.net	2
rediffmail.com	2
snt0-omc1-s49.snt0.hotmail.com	2

Table 6.3: Intersecting SMTP Domains for NPN+RV

NPN+RV	
SMTPServer	**TotalFrequency**
localhost.localdomain	123
mta01.xtra.co.nz	10
na01-bl2-obe.outbound.protection.outlook.com	9
yahoo.com	8
google.com	6
DHE-VE07-1.bps-staff.birmingham.k12.mi.us	4
mail.birmingham.k12.mi.us	4
stcexcpsm04.corp.star	4
1e100.net	3
att.net	3
fep14.mx.upcmail.net	3
HMWEXMB07.AD.HISD.ORG	3
MYMAIL.exeter.edu	3
S0-OTT-X1.nrn.nrcan.gc.ca	3
S0-OTT-XSMTP3.nrcan.gc.ca	3
SSFEXCHEDGE02.srunet.sruad.edu	3
stcexcpsm02.corp.star	3
WCCUSDEXCH01.wccusd.net	3
webmail.exeter.edu	3
ADMIN-IMSS01.HOUSTONISD.ORG	2
co1outboundpool.messaging.microsoft.com	2
correo.ult.edu.cu	2
device.lan	2
edge01.upcmail.net	2
howard.edu	2
localhost.com	2
mail2.wccusd.net	2
rediffmail.com	2
toroondcbmts05-srv.bellnexxia.net	2

Chapter 7

Domain Details

We aimed at performing an Internet-scale study of the current distribution, state and other properties of the existing SMTP servers. The goal is to build a local database of SMTP servers by crawling as many IP addresses as possible. However, we start at a smaller level and use the domains collected from our email datasets. We collect the following information for each of these domains: Domain Name, IP Address, Query Time, Query Date, City, State Name, Country, Zip, Latitude, Longitude, ASN, BGP and State of the SMTP domain. The procedure used to collect these data are as follows:

Using nslookup, nmap and aggregation as earlier, we determined the main SMTP server's state. The state was determined using the Option 3 or the strict closed assumption. The IP address was obtained using the command : dig +short domainName We used a command line tool called geoiplookup to get the following information: stateName, cityName, zipCode, latitude, longitude. And finally we

used the verbose form of whois command from whois.cymru.com to get the following: ASN, BGPPrefix, countryName. The query date and time are saved using the date command.

7.1 Timestamps Visualization

Another interesting factor of the emails that remains unexplored is a time stamps analysis. Each email is associated with a collection of time stamps as appearing in the email header. This takes care of the number of hops in the path of the email as well. We conducted experiments using n-grams from the raw time stamps collection for each email. This produced very good results. But since the overlap of time spans of the different phishing and legitimate datasets was negligible, these results are not dependable.

To visualize the time stamps for each dataset, histograms were created for the frequency of email at each hour of the day. Since the emails traveled through different time zones, two types of histograms are formed. One for the local times and another for the times converted to Coordinated Universal Time (UTC).

Tables 7.1 to 7.8 show the frequencies of emails having the time stamps corresponding to the different hours of the day.

Tables 7.9 to 7.16 show the frequencies of emails having the sent time stamps corresponding to the different hours of the day.

Figure 7.1: Frequency of Legitimate Emails from CSDMC for Local time

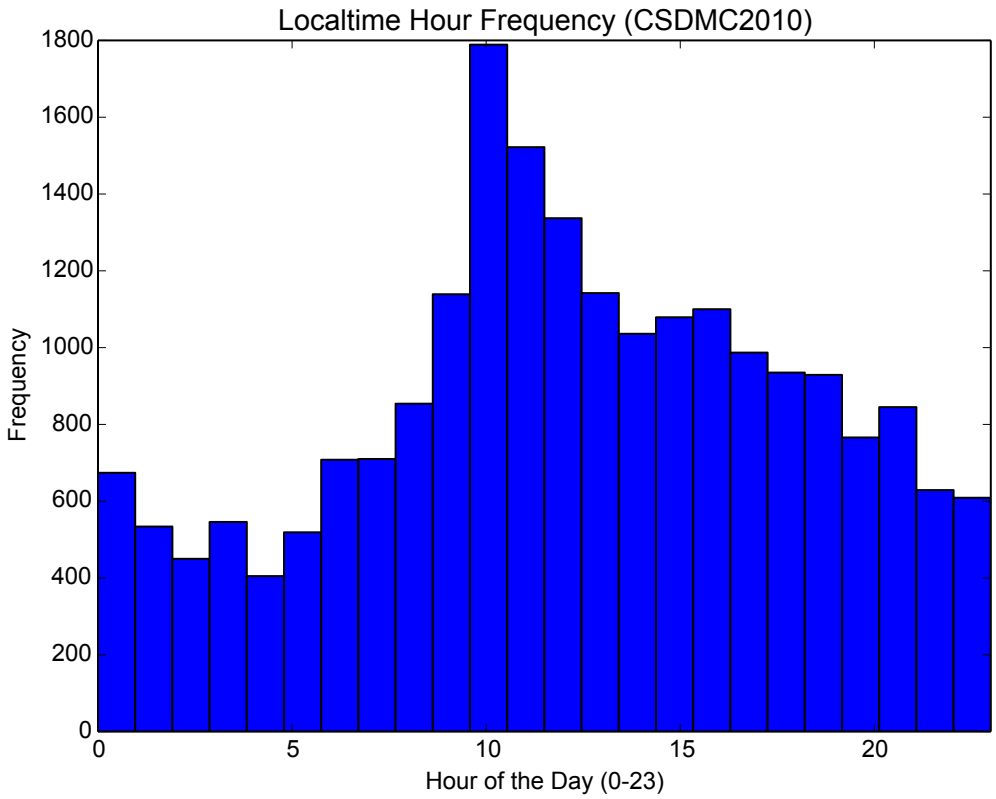


Figure 7.2: Frequency of Legitimate Emails from CSDMC for UTC

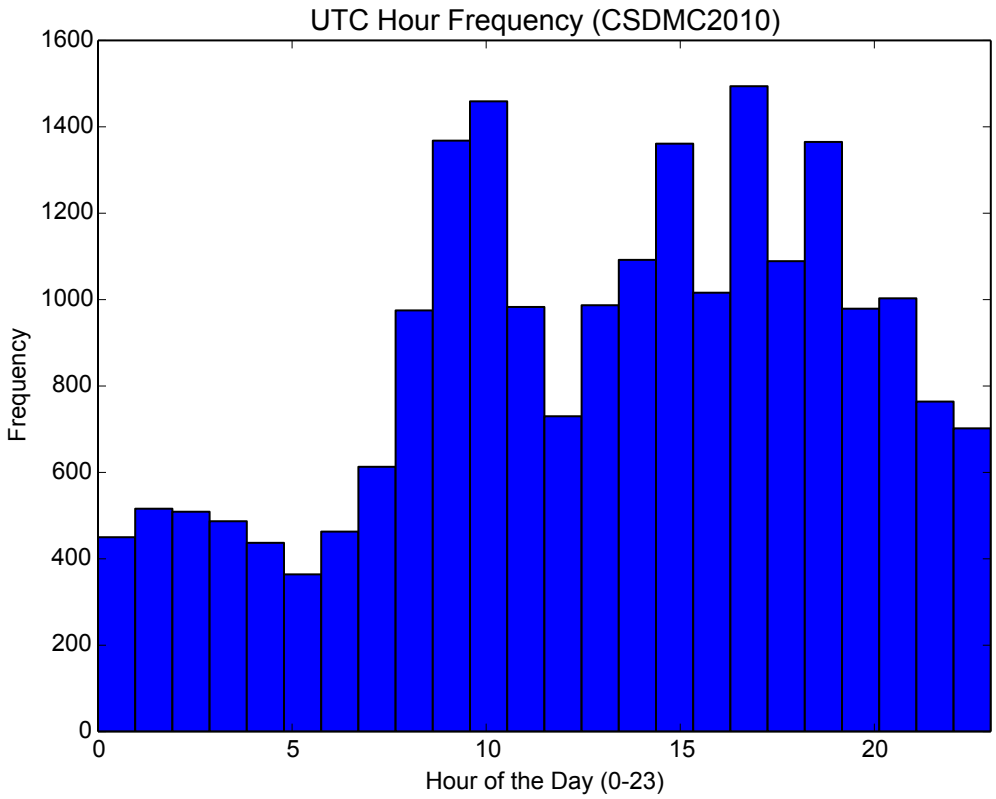


Figure 7.3: Frequency of Legitimate Emails from RVL for Local time

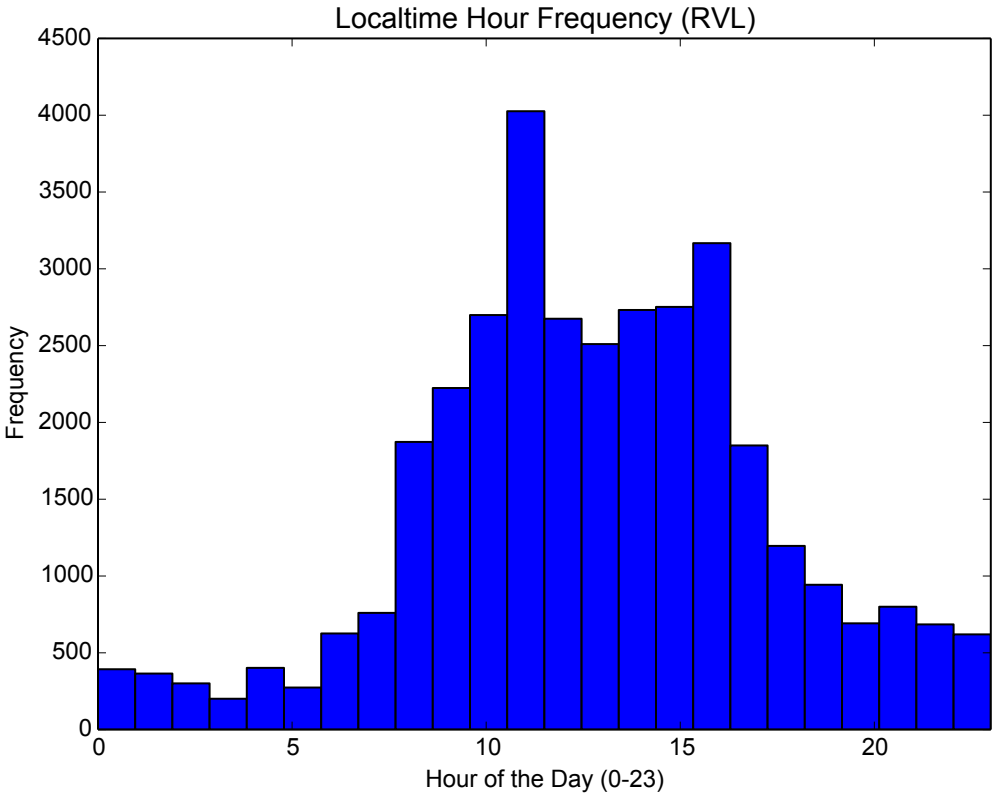


Figure 7.4: Frequency of Legitimate Emails from RVL for UTC

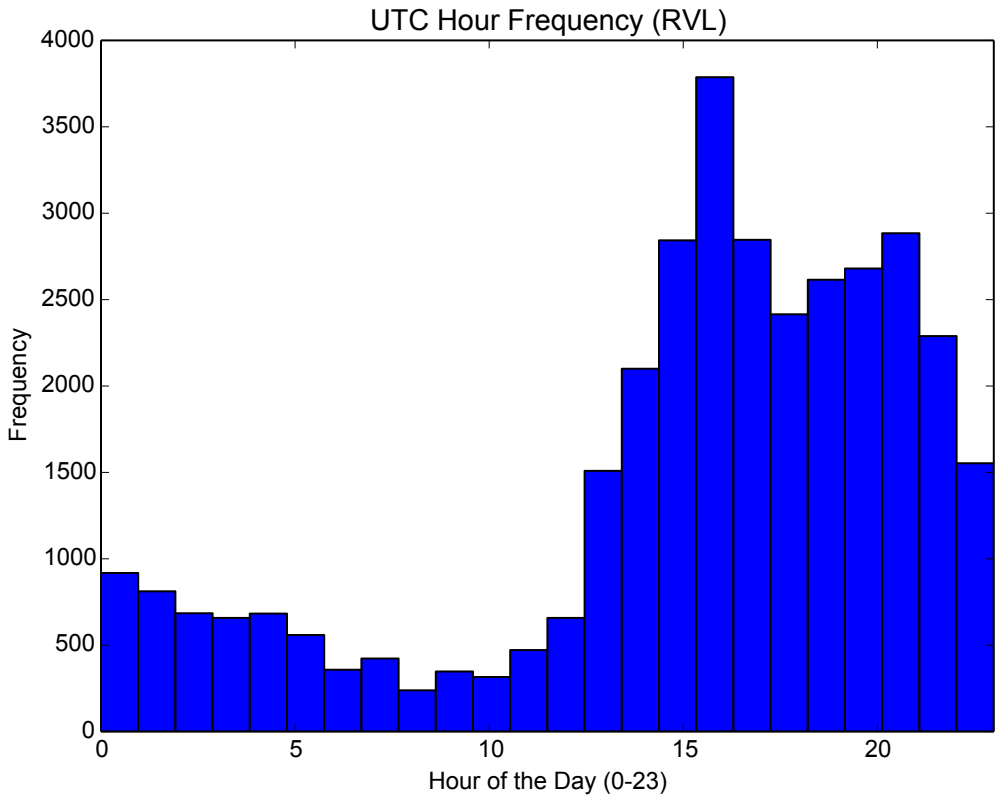


Figure 7.5: Frequency of Phishing Emails from NPN for Local time

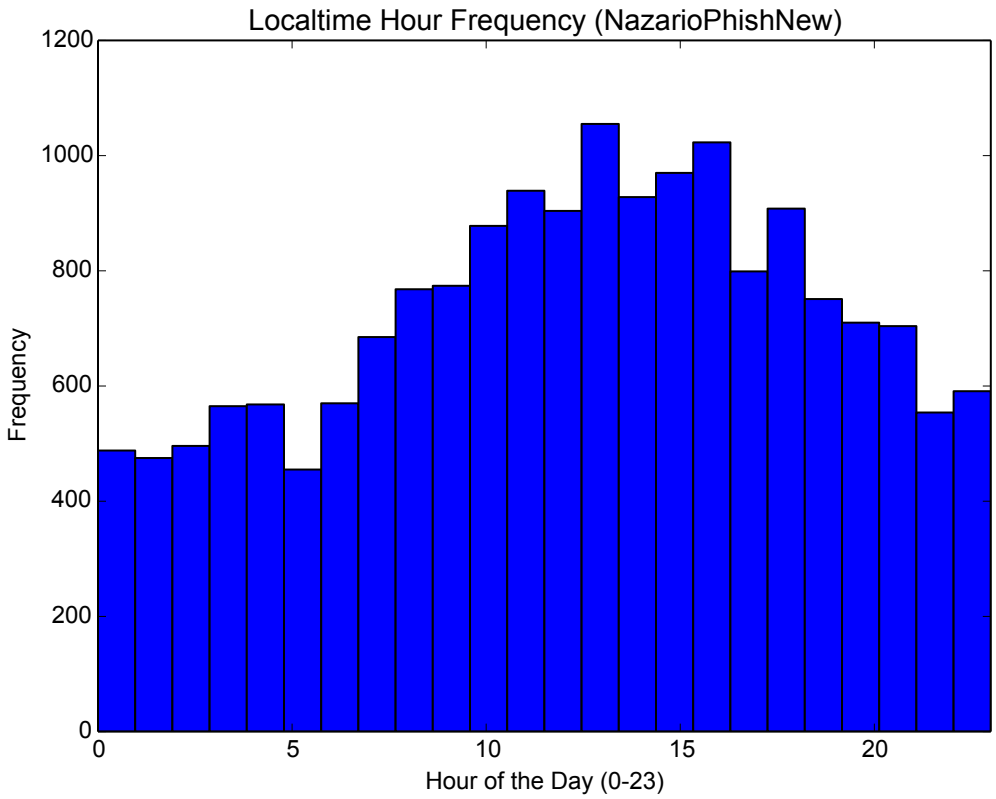


Figure 7.6: Frequency of Phishing Emails from NPN for UTC

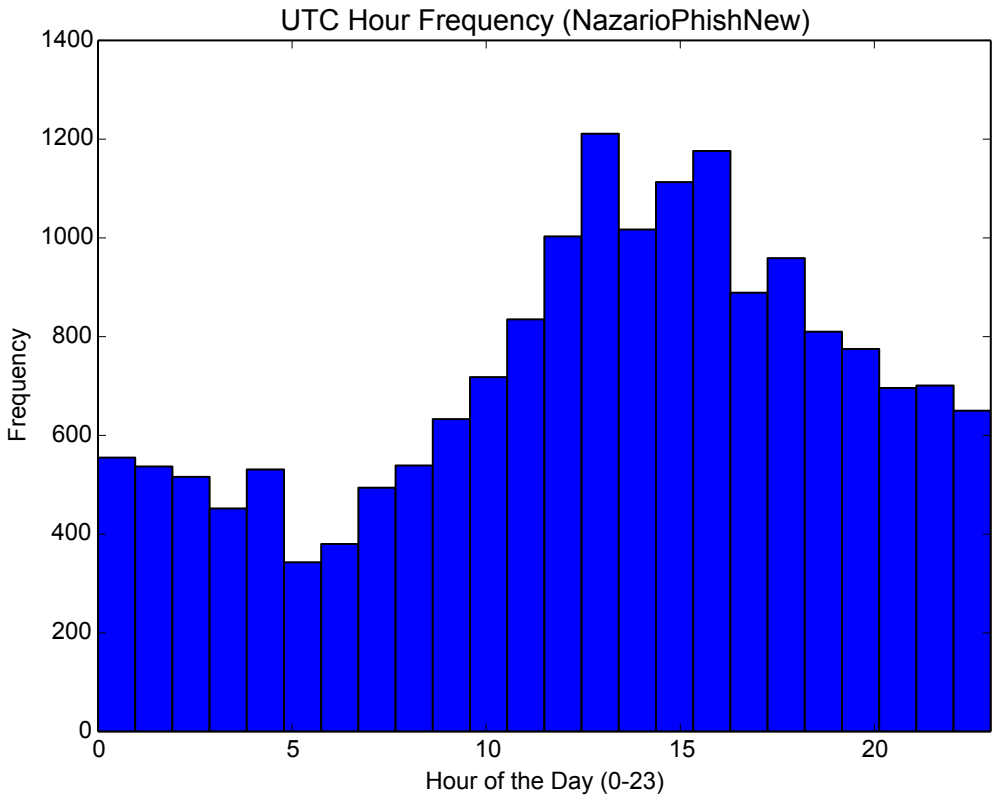


Figure 7.7: Frequency of Phishing Emails from RV for Local time

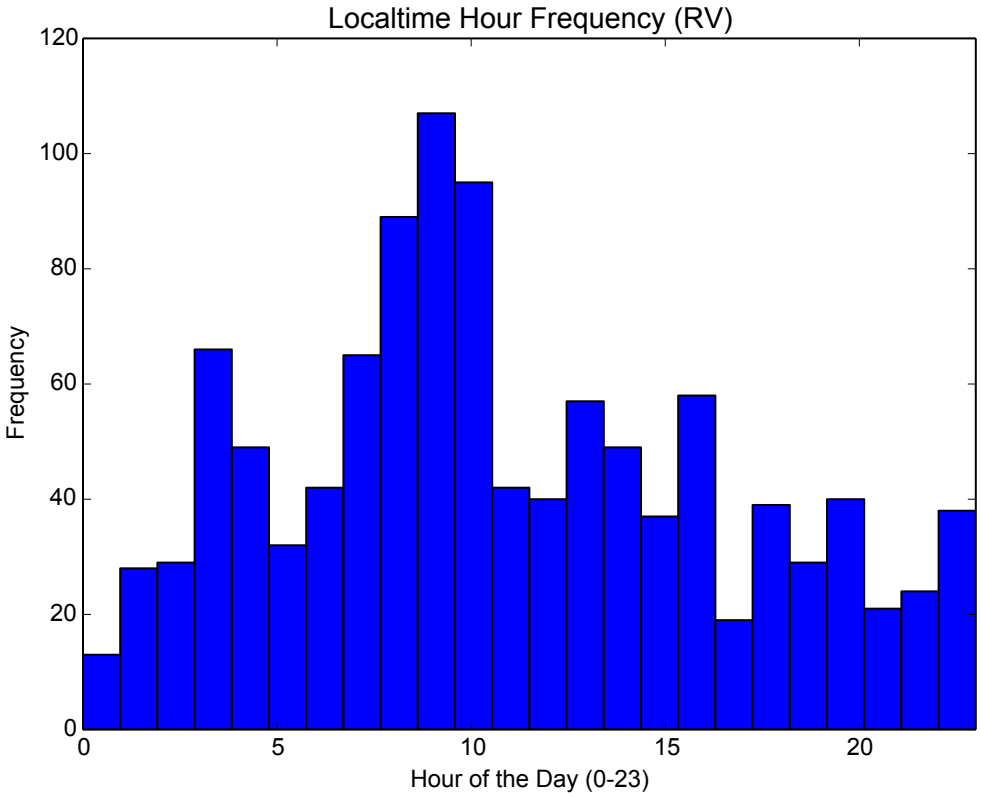


Figure 7.8: Frequency of Phishing Emails from RV for UTC

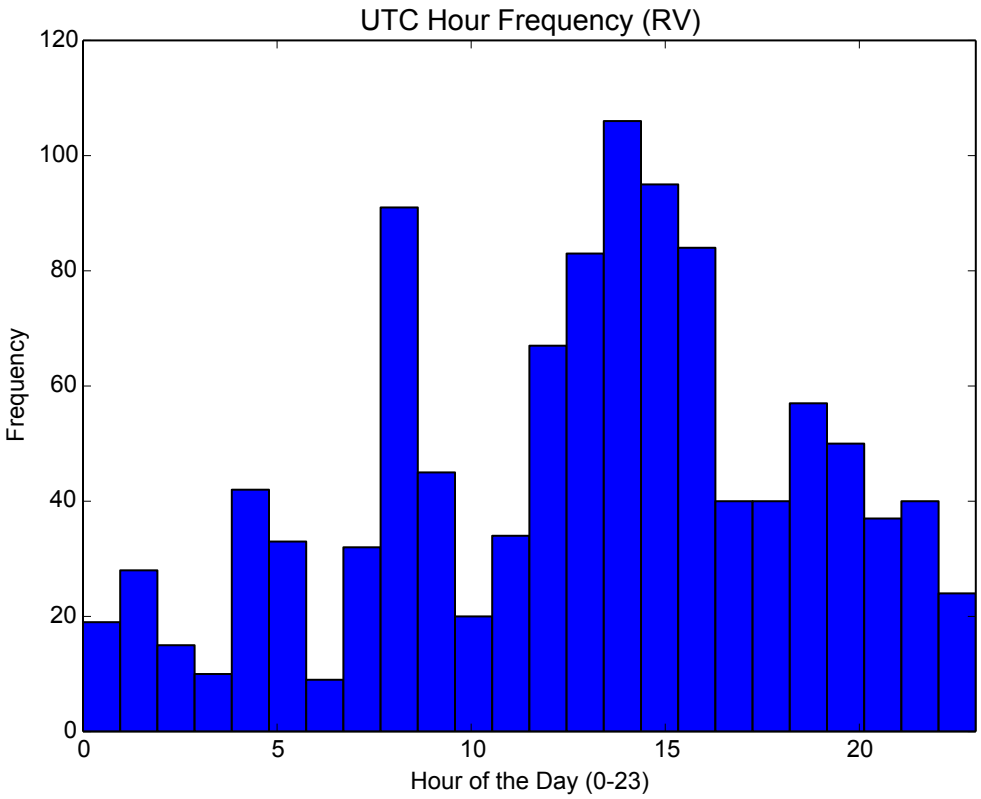


Figure 7.9: Frequency of Legitimate Emails' Sent Times from CSDMC in Local time

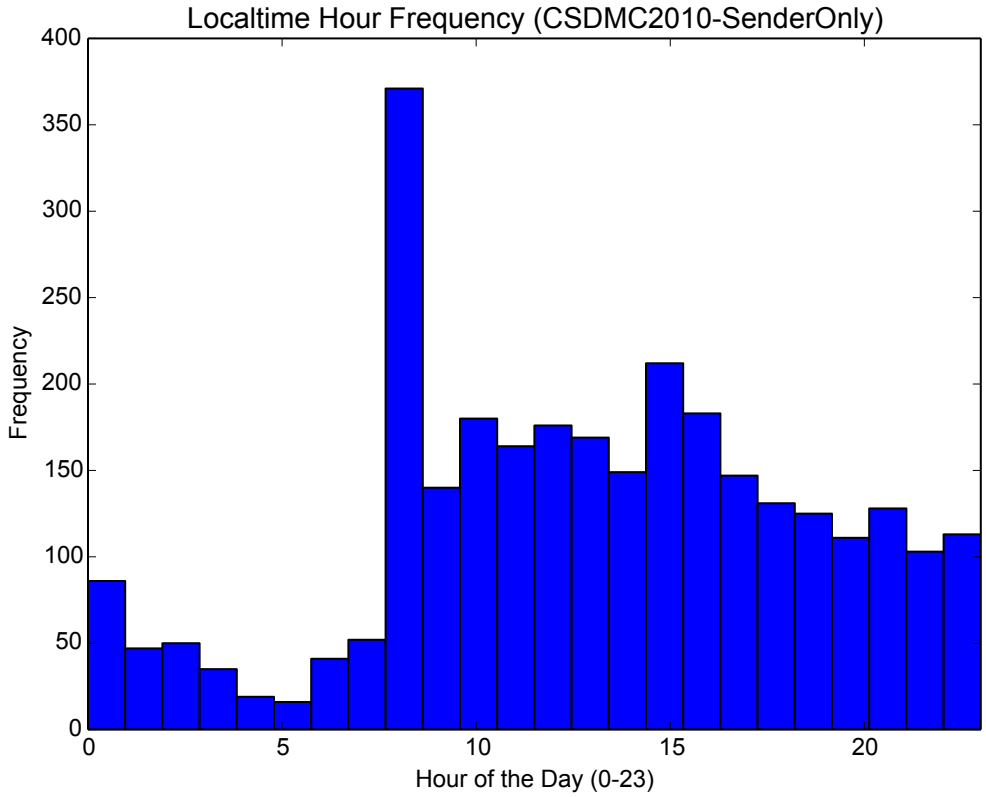


Figure 7.10: Frequency of Legitimate Emails' Sent Times from CSDMC in UTC

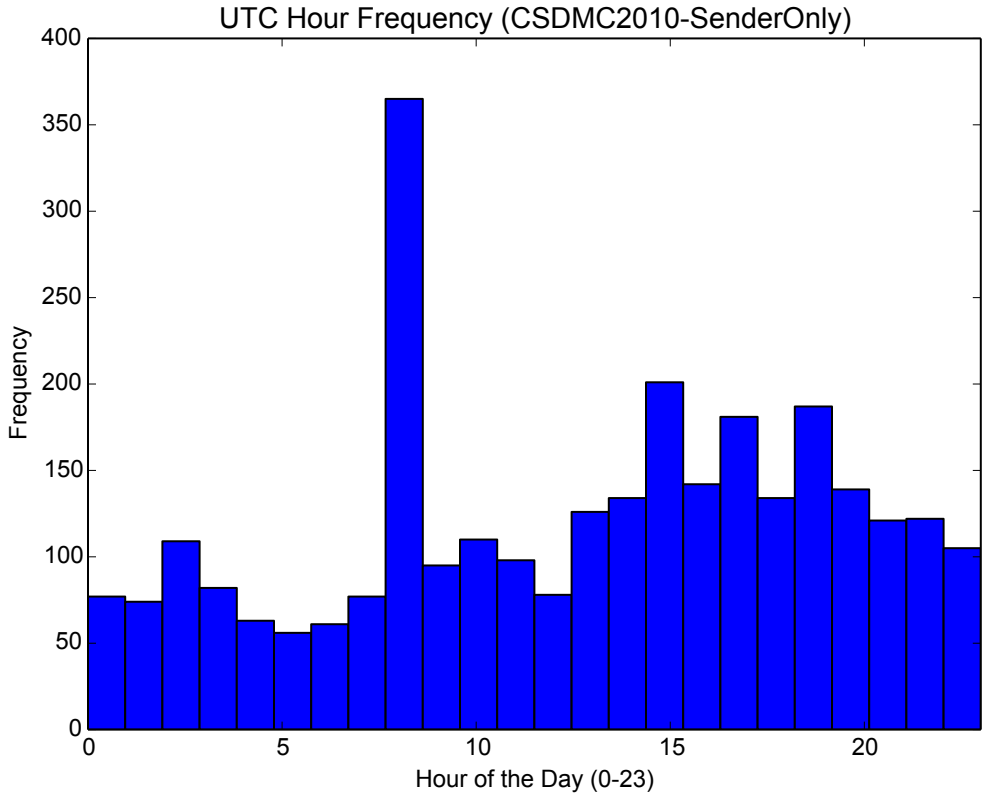


Figure 7.11: Frequency of Legitimate Emails' Sent Times from RVL in Local time

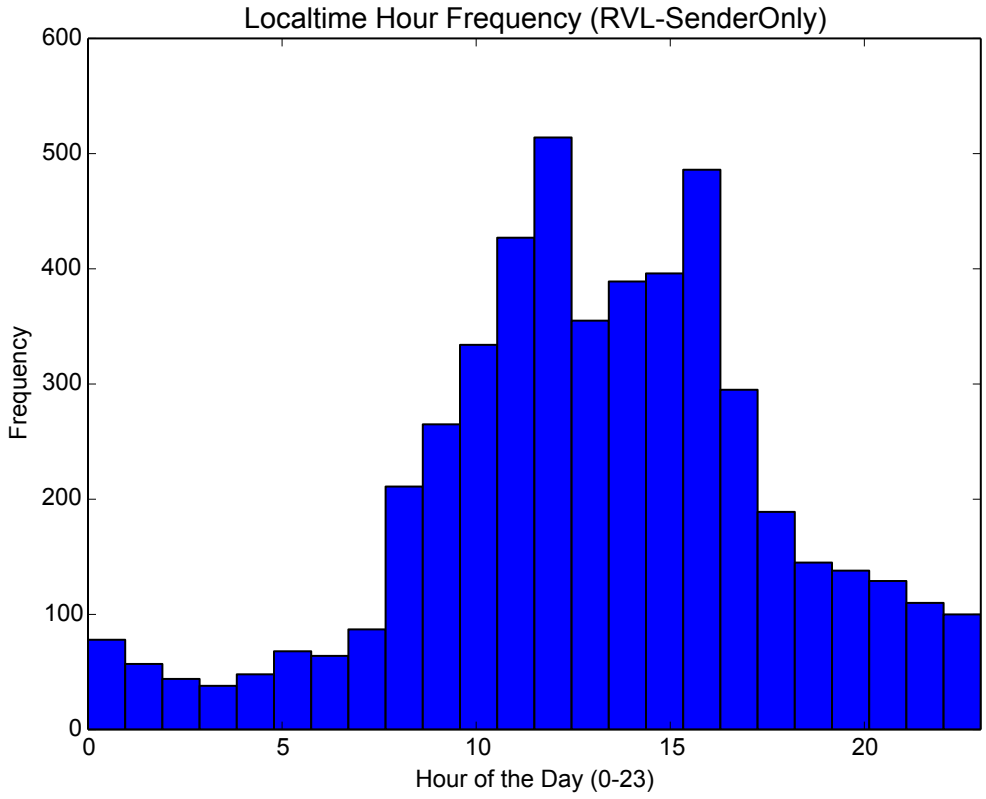


Figure 7.12: Frequency of Legitimate Emails' Sent Times from RVL in UTC

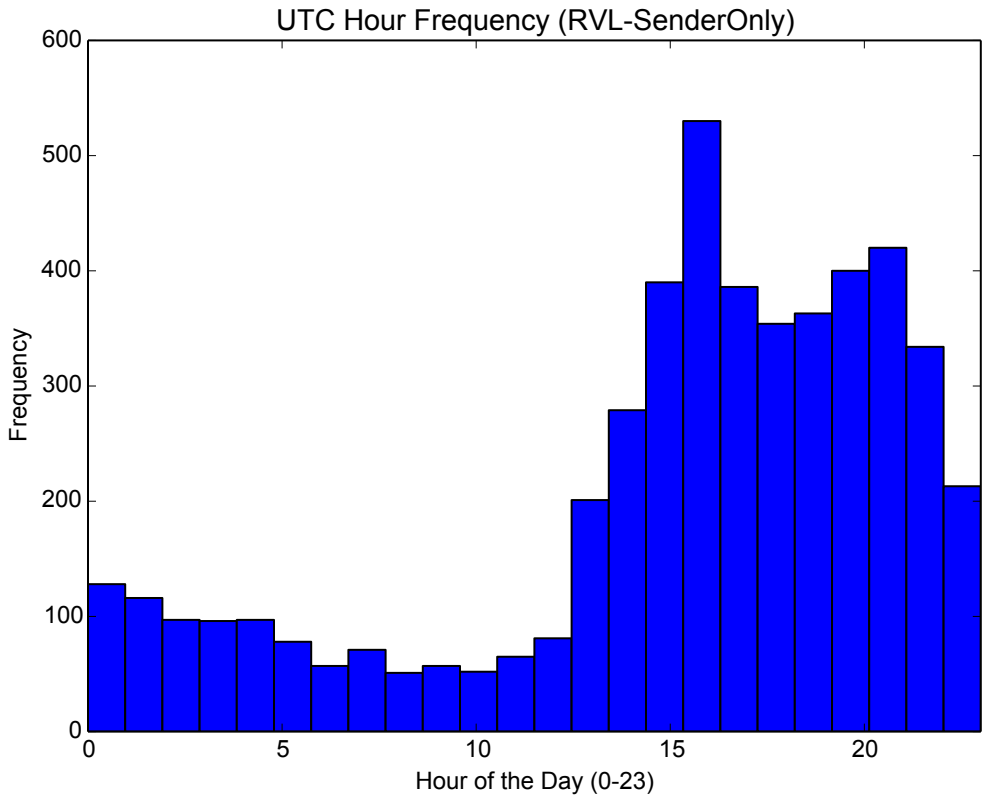


Figure 7.13: Frequency of Phishing Emails' Sent Times from NPN in Local time

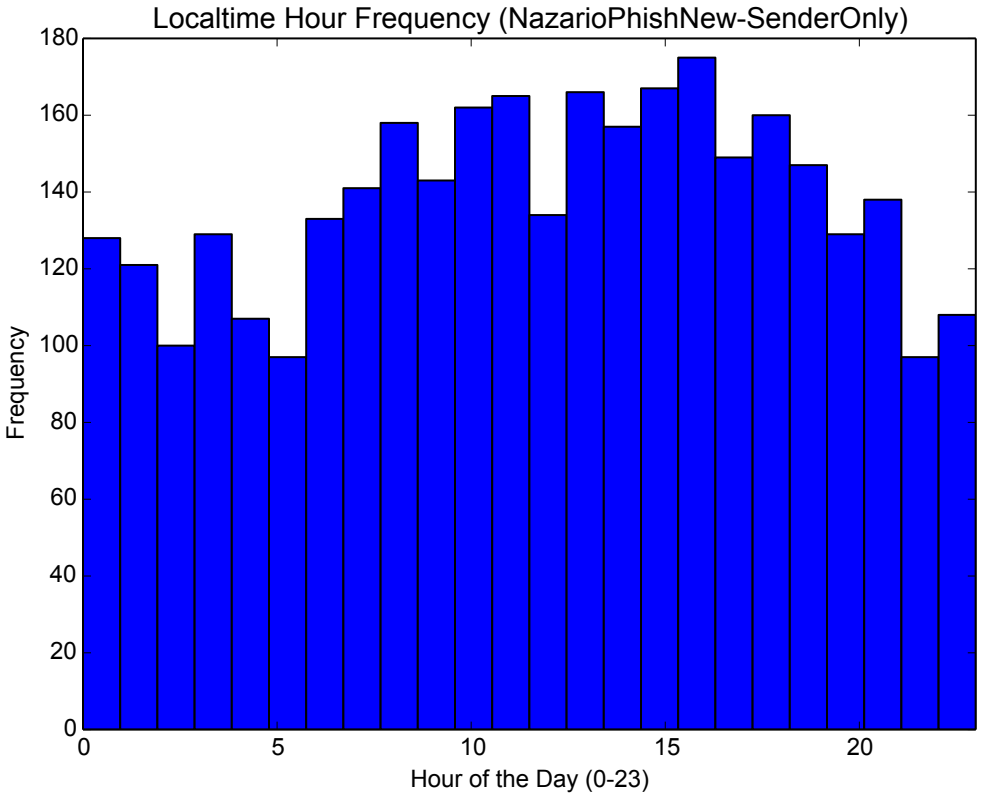


Figure 7.14: Frequency of Phishing Emails' Sent Times from NPN in UTC

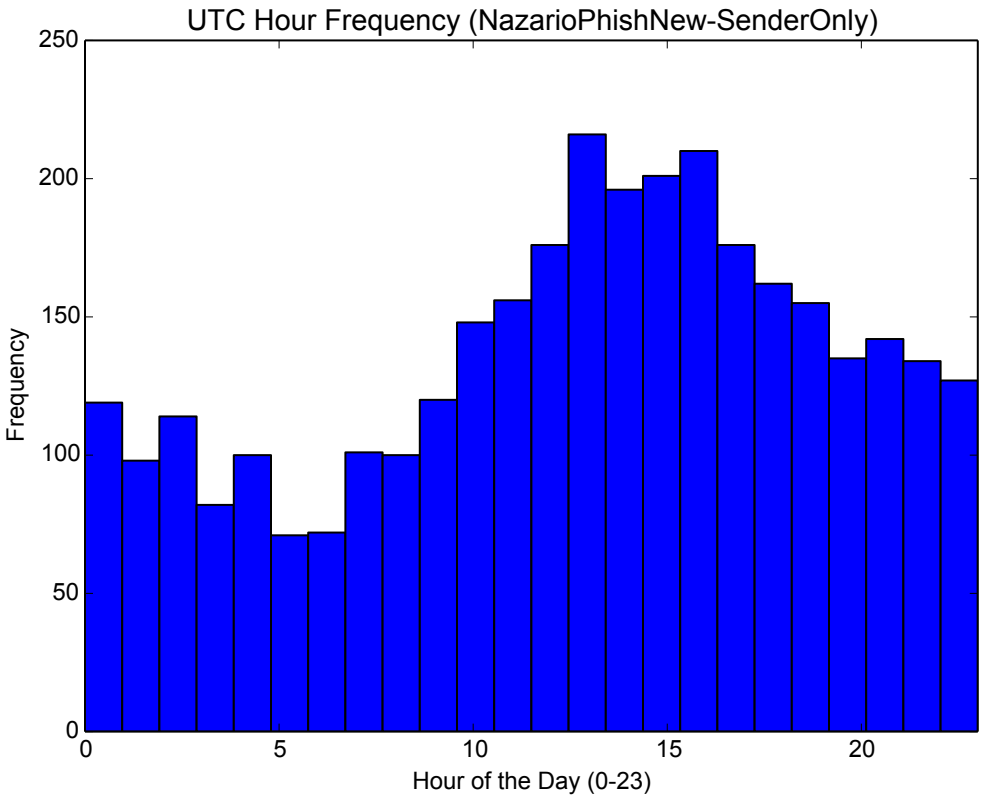


Figure 7.15: Frequency of Phishing Emails' Sent Times from RV in Local time

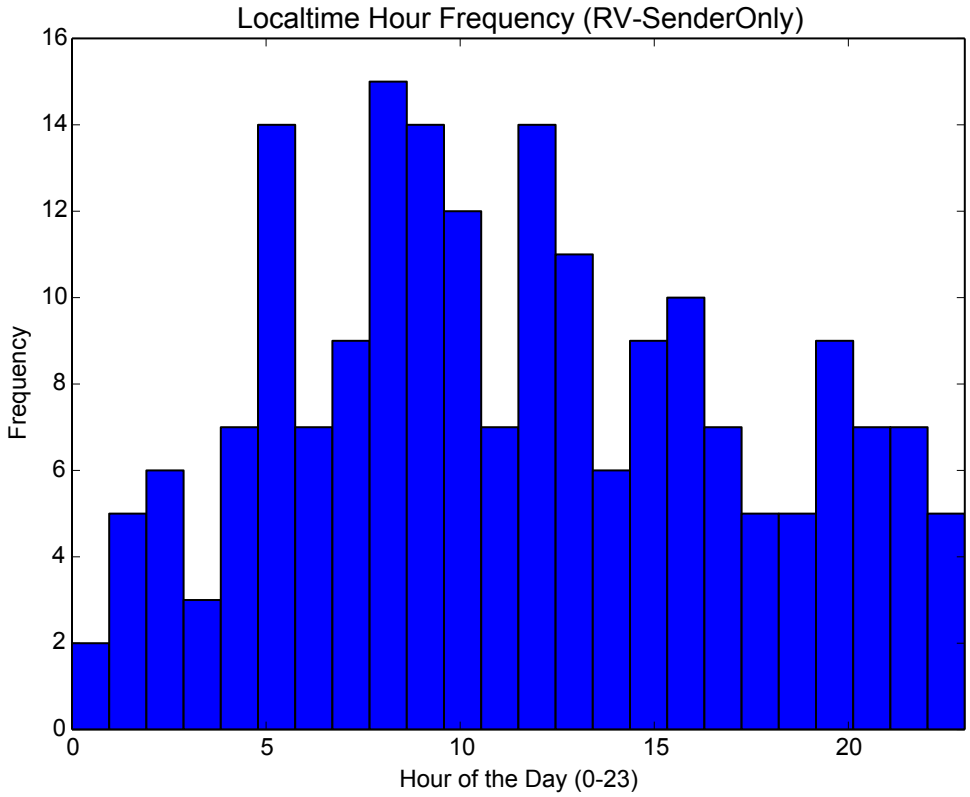


Figure 7.16: Frequency of Phishing Emails' Sent Times from RV in UTC

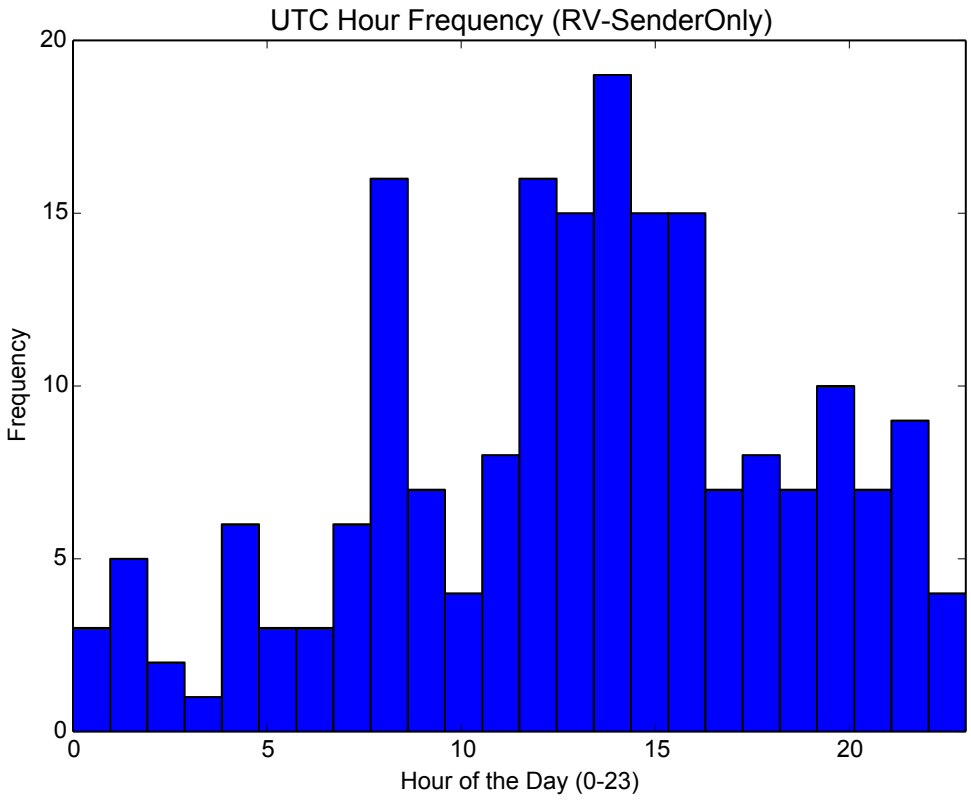


Figure 7.17: Frequency of All Legit Emails' Sent Times in Local time

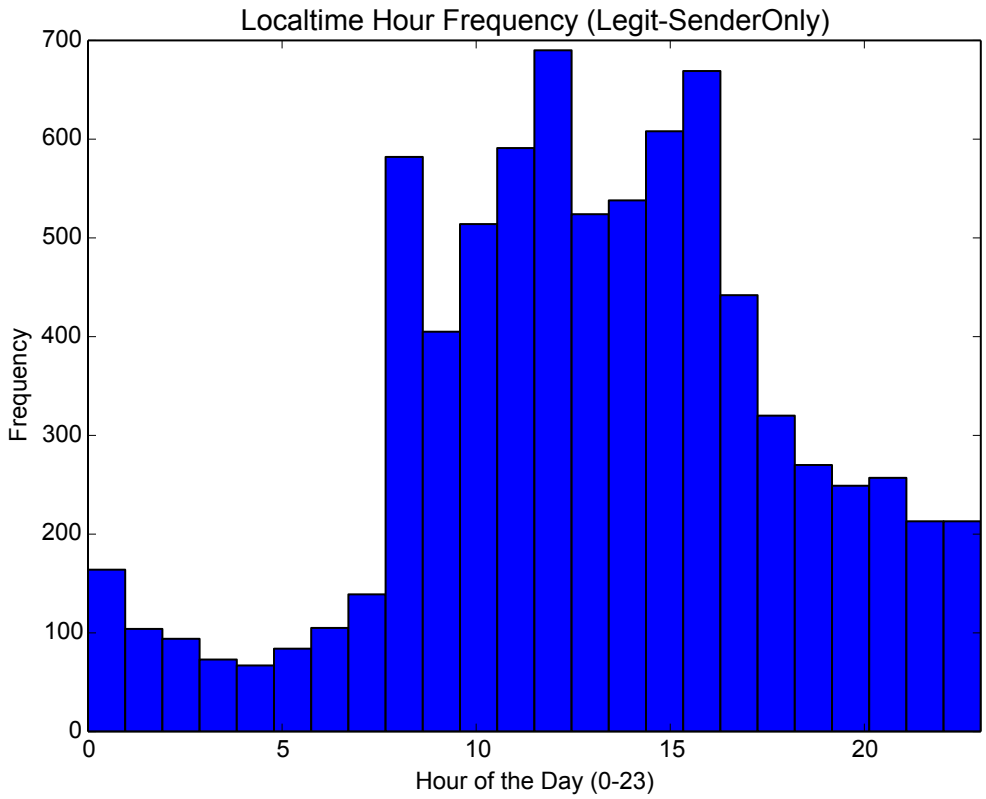
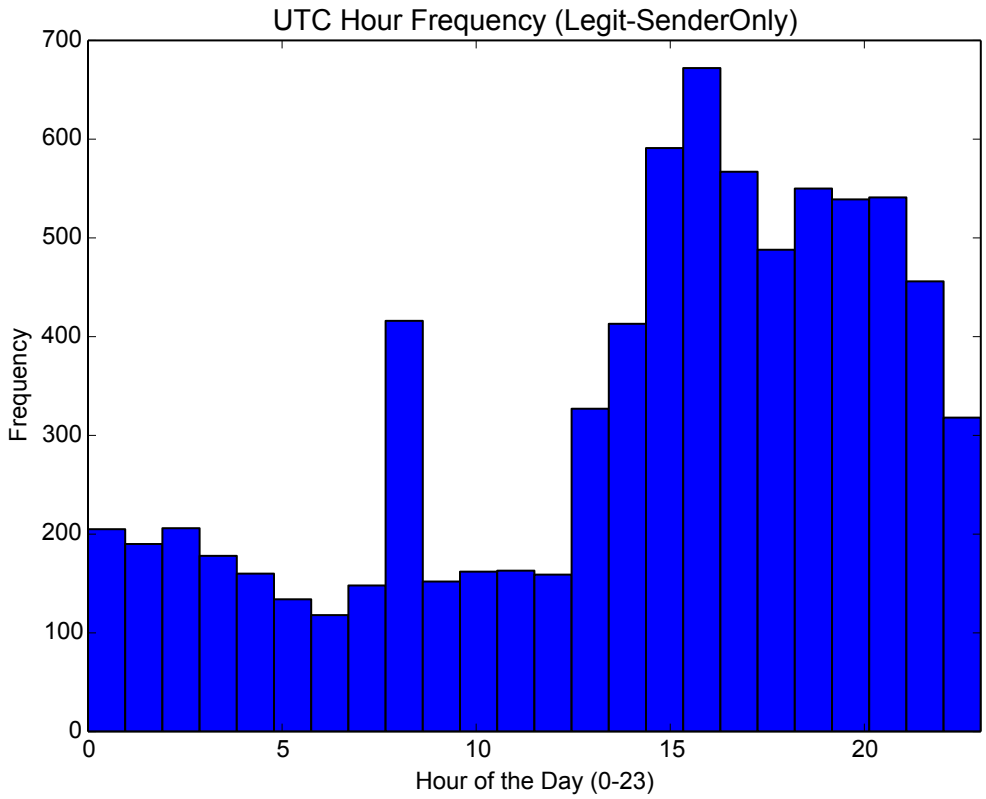


Figure 7.18: Frequency of All Legit Emails' Sent Times in UTC



Tables 7.17 to 7.20 show the frequencies of all phishing and all legit emails having the sent time stamps corresponding to the different hours of the day.

Tables 7.17 to 7.20 show the frequencies of all phishing and all legit emails having the time stamps corresponding to the different hours of the day.

Figure 7.19: Frequency of All Phishing Emails' Sent Times in Local time

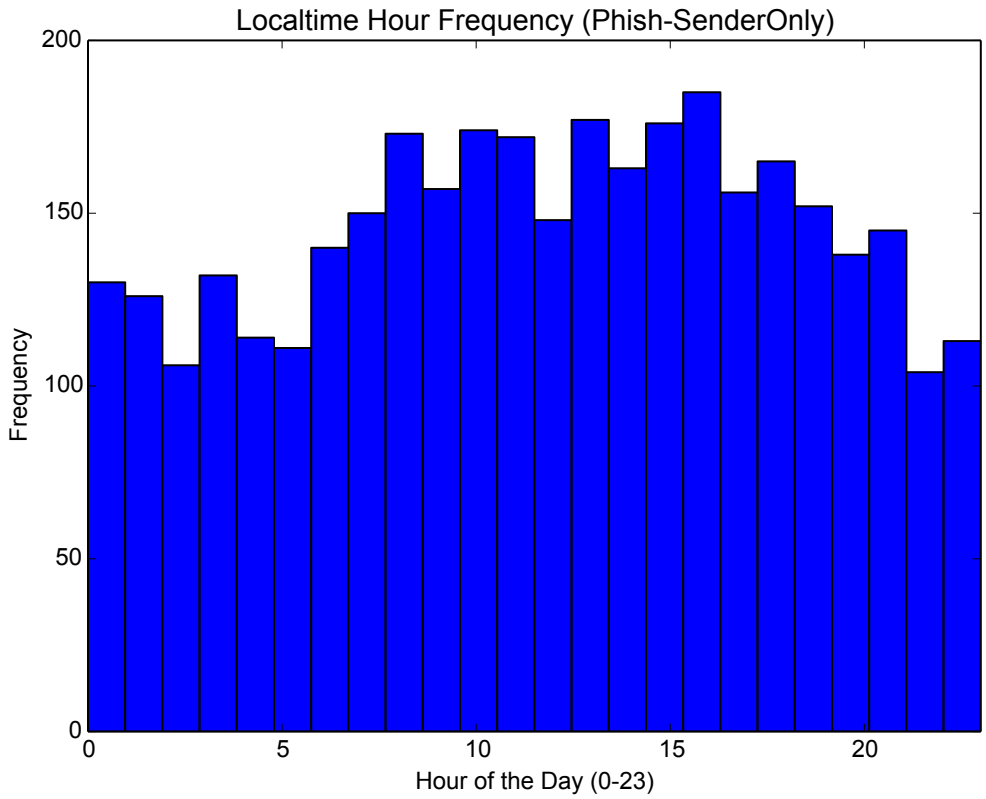


Figure 7.20: Frequency of All Phishing Emails' Sent Times in UTC

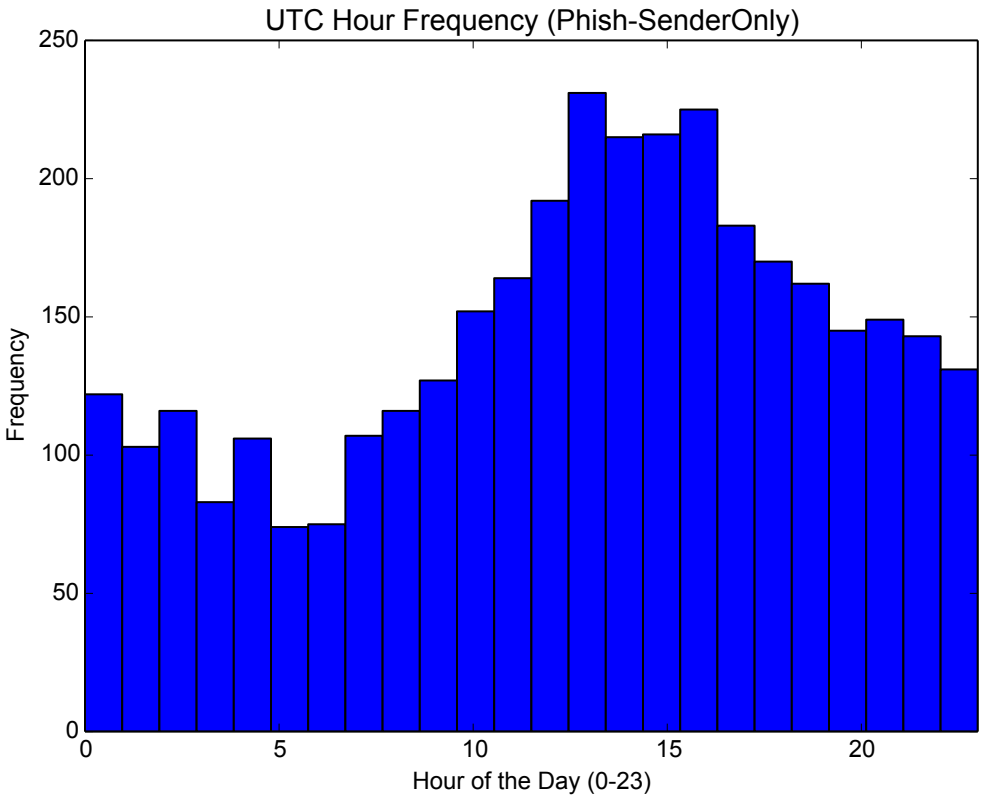


Figure 7.21: Frequency of All Legit Emails for UTC for Local time

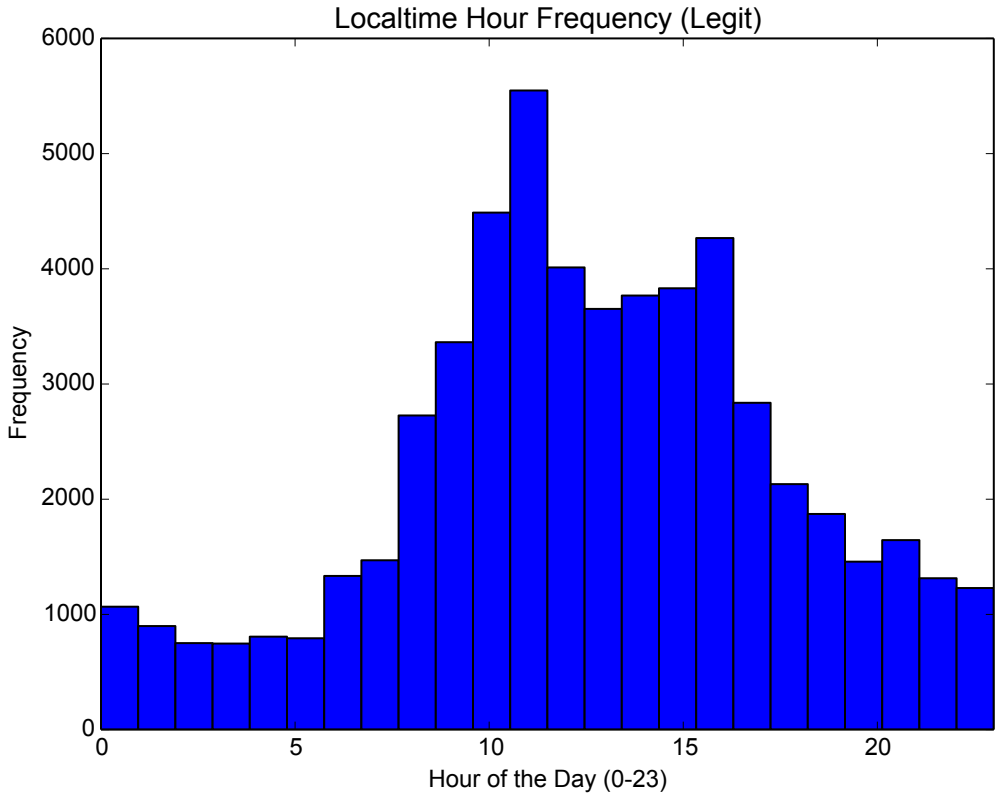


Figure 7.22: Frequency of All Legit Emails for UTC

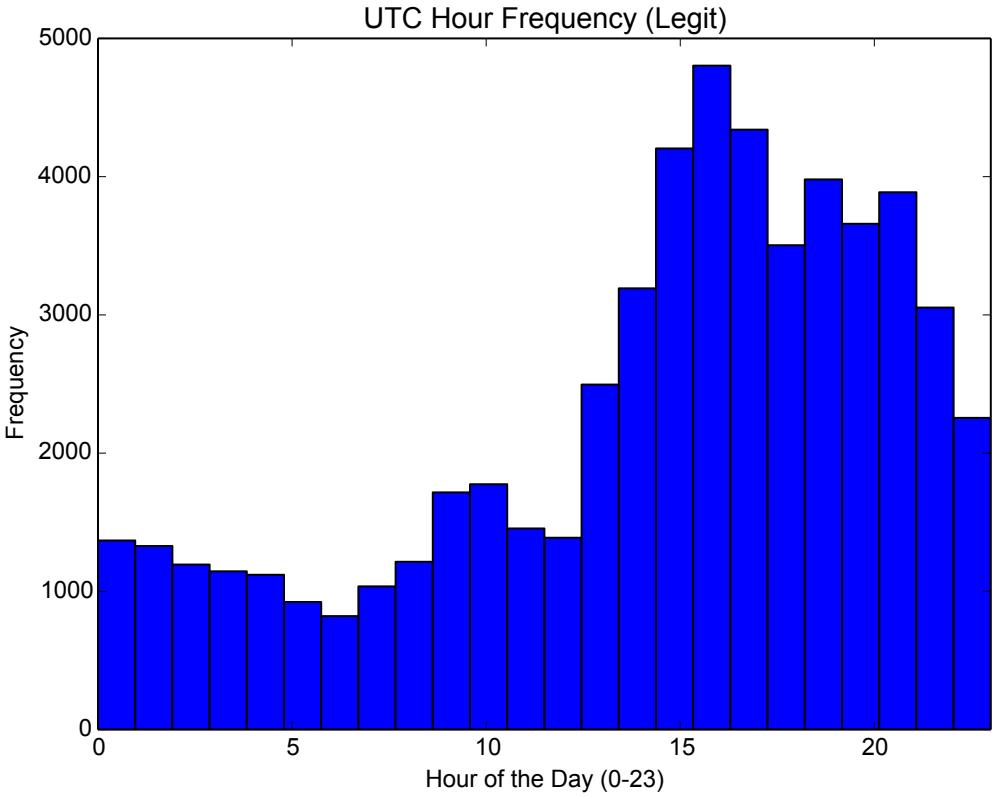


Figure 7.23: Frequency of All Phishing Emails for Local time

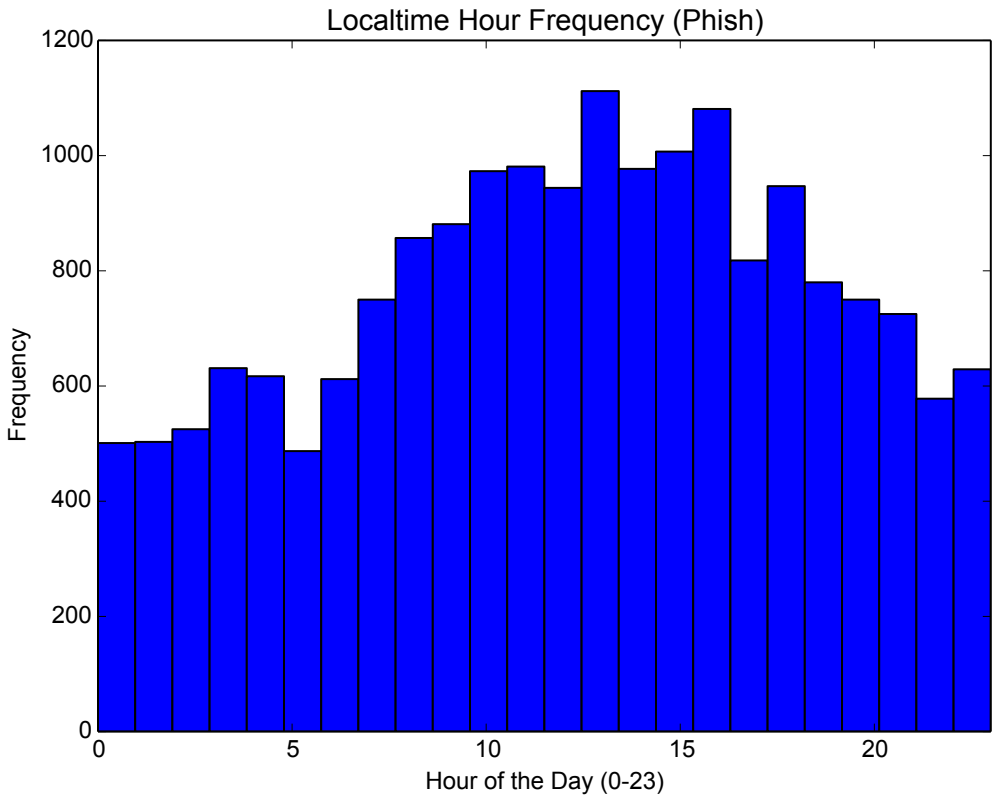
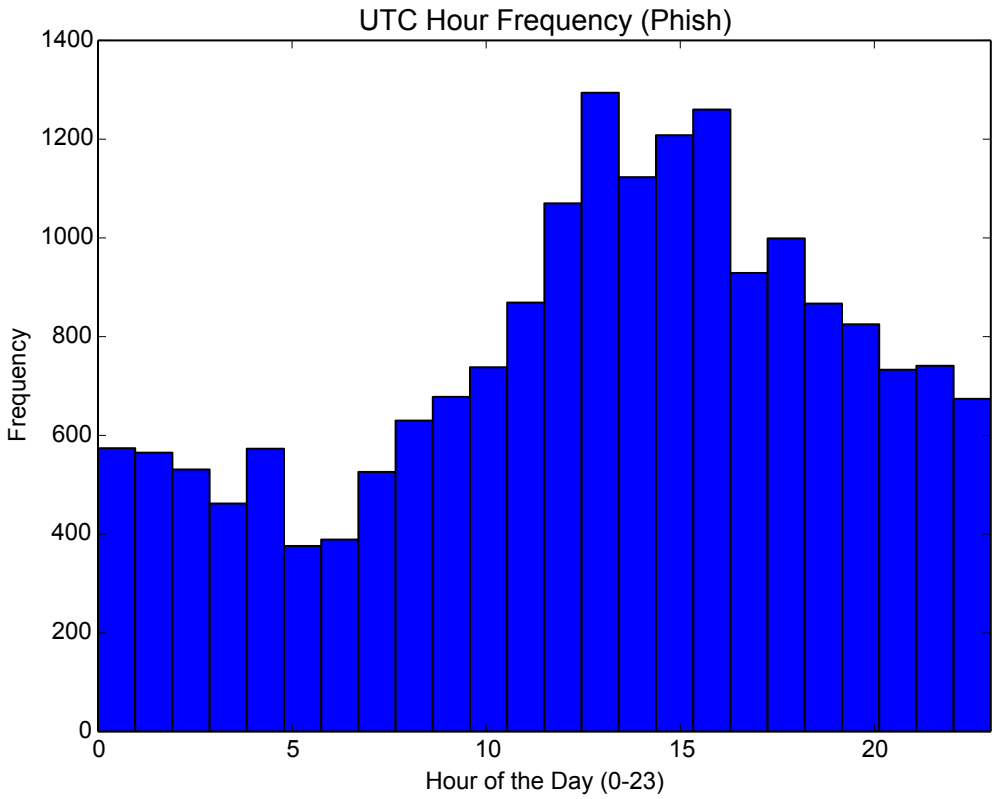


Figure 7.24: Frequency of All Phishing Emails for UTC



Chapter 8

Path Analysis

This chapter deals with the reconstruction of the path taken by the emails. Chapter 5 talks about the supposition that a path once associated with phishing emails is more likely to be associated with other phishing email and we test it using path analysis as described here.

As the email is sent, the first mail server or relay that receives it should have the same domain as the sender. That means the first Received-From domain should match the From field domain. It is checked whether this condition is met by the email or not making it a binary feature. Similarly, for other two binary features it is checked if all 'by' domains are present in the Received-From domains and if the domain matches the accompanying IP in the Received-From field. We also performed preliminary path analysis to check if the email's path is broken. To determine the break in the path we go through the Received-From and 'by' pairs in the email and create the path it traveled using the domains as the vertices or nodes and an

edge between them representing a connection or the absence of an edge representing discontinuity. This results in several continuous features related to the path.

These header fields get added to the email in a bottom-up approach. So the first Received-From field is actually the one in the bottom and the last one is the one at the top. Hence, the path formed also follows the bottom-up approach. An example can be seen in the figure 5.2 where the path can be traced as follows: The email was received from user-119ac86.biz.mindspring.com by maynard.mail.mindspring.net, from maynard.mail.mindspring.net by xent.com, from lair.xent.com by xent.com and so on.

8.1 Subroutines

We tried various ways of extracting from the header information about the path that an email has taken from the sender to the recipient. Four different checks were performed and the result converted to features for each email to determine the legitimacy of the path taken. Before creating the features, we performed the **Comprehensive Extraction of Email Header Information** from all the emails. This included the extraction of all the fields and data that we found relevant for the purpose of email header analysis. Domains from all the header fields, ESMTPIDs, Message-IDs, X-Mailer information, X-Spam information, Timestamps, Received-from and By pairs, would be a good representation of the data extracted. This was done to facilitate any further header analysis experiments involving the data from the headers. The four different checks performed are in the form of the following

subroutines:

8.1.1 From Received-From Mismatch

Ideally, the From field domain of an email should match the domain of the first Received-From field, i.e., the Received-From closest to the sender. This subroutine checks if this condition is satisfied. If yes, it returns 0 otherwise 1 which is also the value of the binary feature 'From Received-From Mismatch'.

8.1.2 All By in Received-From

The domains following the 'by' field in the emails are the ones which have received the emails. Since they are receiving the emails, it is proper that they become the next domain from which another domain will receive the email. So, a 'by' domain must also be in the Received-From domain. This subroutine checks if all the 'by' fields are present in the Received-From fields. If yes, it returns 0 otherwise 1 which is also the value of the binary feature 'All By in Received-From'.

8.1.3 Claiming Domain Different from Actual Domain

Many domains also provide an IP along with them in the email headers. In case of obfuscation the email might be claiming to be from a domain but actually belong to another. The subroutine determines if that is the case by checking if the domain from the given IP matches the given domain. If yes, it returns 0 otherwise 1 which is also

the value of the binary feature 'Claiming Domain Different from Actual Domain'.

8.1.4 Path Broken

For this check, we try to recreate the complete chain of edges formed by the from-by pairs in the email header. We keep track of connected edges where the Received-From match the previous 'by', the disconnected edges where only a From-By pair is available without any connection to the next edge, and the orphan nodes which are the domains that are not a part of any From-By edge pairs. We calculate the total path length, the connected edges count and ratio, the disconnected edges count and ratio from the path analysis. Also, we derive the number of breaks and the distance of the first break from the sender. This results in several continuous features as follows: TotalConnectedEdges, TotalDisjointEdges, TotalOrphanNodes, PathLength, ConnectedRatio, DisjointRatio, OrphanRatio, BreakPosition1, BreakPosition2 till maximum number of breaks.

To explain this with a simple example consider the following chain: a-b, b-c, c-d, e-f, g, h-i, where a, b, c .. i are the domains. Here we can see that the connected edges are a-b, b-c and c-d. The disconnected edges are e-f and h-i and the orphan node is g.

8.2 Results

Table 8.1 summarizes the results of the weka experiments on the combined dataset CSDMCRVL+NPNRV for the combined features of path analysis. These results show that the best result is obtained for Random Forest classifier with 93% TPR and 13.3% FPR.

Figure 8.1: The received Header Fields of an Email

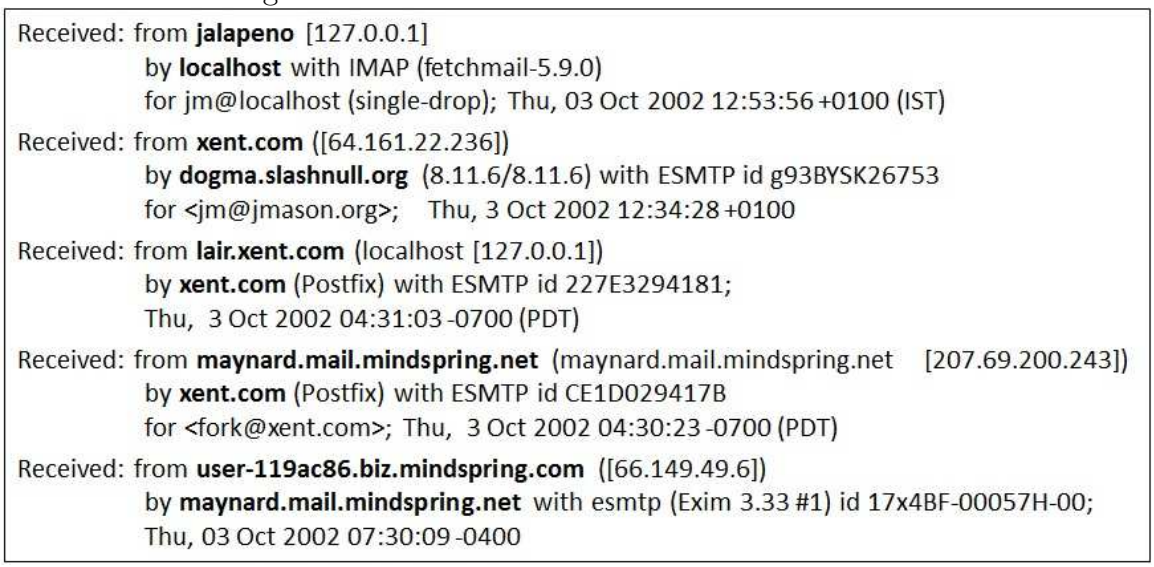


Table 8.1: Results for combined features of path analysis

PathAnalysisCombinedCSDMCRVL+NPNRV							
Class	**Classifier**	**TPR**	**FPR**	**Precision**	**Recall**	**FMeasure**	**ROC**
legit	**ABoost**	0.819	0.329	0.846	0.819	0.833	0.798
phishing	**ABoost**	0.671	0.181	0.627	0.671	0.648	0.798
Weighted	**ABoost**	0.773	0.283	0.778	0.773	0.775	0.798
legit	**AttSel**	0.902	0.355	0.849	0.902	0.875	0.889
phishing	**AttSel**	0.645	0.098	0.749	0.645	0.693	0.889
Weighted	**AttSel**	0.822	0.275	0.818	0.822	0.818	0.889
legit	**Bagging**	0.981	0.196	0.917	0.981	0.948	0.976
phishing	**Bagging**	0.804	0.019	0.952	0.804	0.872	0.976
Weighted	**Bagging**	0.926	0.14	0.928	0.926	0.924	0.976
legit	**BLR**	0.999	0.981	0.692	0.999	0.818	0.509
phishing	**BLR**	0.019	0.001	0.907	0.019	0.037	0.509
Weighted	**BLR**	0.694	0.676	0.759	0.694	0.575	0.509
legit	**J48**	0.982	0.193	0.919	0.982	0.949	0.971
phishing	**J48**	0.807	0.018	0.952	0.807	0.874	0.971
Weighted	**J48**	0.927	0.138	0.929	0.927	0.926	0.971
legit	**NB**	0.478	0.129	0.891	0.478	0.622	0.768
phishing	**NB**	0.871	0.522	0.43	0.871	0.576	0.768
Weighted	**NB**	0.6	0.251	0.748	0.6	0.608	0.768
legit	**RF**	0.983	0.186	0.921	0.983	0.951	0.981
phishing	**RF**	0.814	0.017	0.955	0.814	0.879	0.981
Weighted	**RF**	0.93	0.133	0.932	0.93	0.929	0.981
legit	**SMO**	0.969	0.782	0.733	0.969	0.834	0.593
phishing	**SMO**	0.218	0.031	0.759	0.218	0.339	0.593
Weighted	**SMO**	0.735	0.548	0.741	0.735	0.68	0.593

Chapter 9

Related Work

Since it is a much-employed security threat, automatic detection of phishing emails has attracted significant attention of researchers over the last decade or so. We highlight and compare our work with respect to best previous related research.

9.1 Phish-IDetector

One attempt to phishing emails classification was made by PILFER [16]. This paper lists 10 features, both binary as well as continuous numeric ones designed to highlight user-targeted deception in electronic communication. These features were mainly based on URL information like: IP based URLs, Age of linked to domain names, Nonmatching URLs, Number of dots (in the URLs) etc. There were also some other feature that considered if the emails were in HTML format, the site of redirection, output from spam filter, etc. Applying machine learning on these extracted features

via random forest classifier with 10-fold cross validation they could correctly classify over 96% of the emails. Their false-positive rate was of the order of 0.001%. The datasets they used were the same as ours but being an older version, the number of emails were significantly less in their experiments. They had a total of 6950 non-phishing emails from SpamAssassin [3] dataset and only 860 phishing emails from the Nazario [30] dataset.

The structural features of the emails such as 'spoofing of online banks and retailers', 'link in the text is different from the destination', 'using IP addresses instead of URLs', etc. were studied by [10] and these features were selected using the simulated annealing algorithm. They found that these structural features when combined with one class support vector machine (SVM), could be used to efficiently classify the phishing emails before it reaches the users inbox, essentially reducing human exposure. They claimed a 100% precision and recall. However their data set was small consisting of only 400 emails in total. Half of them were phishing and the other half were legitimate emails. The phishing emails were collected over a period of 6 months. And the legitimate ones were gathered from (i) postings on newsgroups, bulletin boards, and from other users inbox and (ii) from 8 different volunteers who provided emails sent to them from legitimate business organizations such as credit card statements, online purchase receipts from Amazon, etc.

Some methods utilize the confidence weighted linear classifiers like [5] but it is only applied to the email body or text unlike our approach. They use the contents of the emails (word stems) as features without applying any heuristic based phishing specific features and the best accuracy obtained was 99.77% which is 99.99% in our

case. They obtained a false positive rate of less than 1%. In simple terms, word stem refers to the simple form or the root of a word. For example, run is the stem for running. They represented each email document as a vector of stemmed words which is commonly known as 'bag of words' representation. For the phishing dataset they used the Nazario corpus and for the legitimate emails they used the SpamAssassin ham corpus.

A hybrid feature selection approach based on combination of content-based and behavior-based features was put forward by [20]. It could mine the attacker behavior based on email header and utilized the Message-ID tags of emails to do so. The authors analyzed the Message-ID tag and sender email to form a feature called Domain_sender. It is a binary feature that represents the similarity of domain name extracted from email sender with domain Message-ID. If it is similar, the email is considered legitimate and the value is set to 0 otherwise 1. This method of hybrid features selections were able to achieve 96% accuracy rate and 4% false positives rate. For the phishing emails, they used the same data set as ours but for the legitimate ones, they only used the easy ham directory of the SpamAssassin corpus which contained only 2364 ham emails.

Another paper that goes deep into email header analysis is [32]. It studies the Message-ID field minutely and explains each part that constitute it. Message-ID generation is discussed in details and the uniqueness of this field is established. The author shows that spoofing of this field is tough and may not be possible for every phisher unless he has sound technical knowledge in this field. Hence, the author suggests that Message-ID could be used to find out about the source of the email

which could be useful in forensic analysis.

It is also worth mentioning the work done in the field of phishing emails detection using the information in the email header, links and body by [41]. They included natural language processing tools and techniques along with contextual information from a user's mailbox in their email header and body analysis. For the body of the email they calculate Textscore using lexical analysis, part-of-speech (POS) tagging,etc. along with verb analysis of action words. They also calculate the Contextscore considering the email as a vector of TFIDF. For header analysis they look at the From, Delivered-to and Received-From fields. And in case of link analysis, they consider the length of the domains in the url and employ google search to ensure authenticity of the domain. Their method was able to correctly classify 98% of the 2000 phishing emails and 99.3% of the 1000 legitimate emails.

9.2 Header-Domain Analysis

Please refer to Table 9.1.

Table 9.1: Related Work for Header Domain Analysis

Classifier	Summary	Features	Results	Datasets
Weka: C4.5 Decision Tree,Support Vector Machine, Multilayer Perception, Nave Bayes, Bayesian Network and Random Forest (RF).[2]	Identify potential header features for spam filtering using machine learning classifiers.	1. Received field (Hops, Span time, Domain add, Date, time, IP Add legality) 2. Sender add 3. No. of receivers 4. Reception Date 5. X-Mailer 6. Missing/ malformed Msg-ID 7. Subject	Best: RF classifier Avg. acc: 98.5% Precision: 98.4% Recall: 98.5% F-Measure: 98.5% ROC area: 99%.	CEAS2008 live spam challenge lab corpus (26180 spam and 6523 ham) CSDMC 2010 (1378 spam, 2949 ham)
No classification performed.[43]	Analyzed sender and receiver field information to identify spam.	1. Sender add validity 2. Receiver add (To, CC, BCC)	No classification performed.	3,417 mails from Taiwan's ISP.
Random Forest (RF) classifier.[21]	Presents an Intelligent Hybrid Spam Filtering Framework (IHSFF). Can identify spam based on email header.	1. Originator field (From) 2. Destination field (To, CC, BCC) 3. X-Mailer 4. Sender server IP add 5. Subject	Best: RF Accuracy: 96.74% Precision: 93.53% Recall: 92.99% F-Measure: 93.26%.	From a Chinese website. Dataset 1: 33,209 samples Dataset 2: 21,725 email headers.
RF, C4.5 DT (J48), Voting Feature Intervals, Random Tree, REPTree, Bayesian Network and Naive Bayes.[35]	Studies information in the email header and evaluate those features with several machine-learning classifiers.	1. From field 2. To and CC 3. Received 4. Message-ID 5. Return-Path 6. Reply-To 7. In Reply-To 8. Error-To 9. Sender 10. Reference	Best: RF Accuracy: 99.27%, Precision: 99.40% Recall: 99.50% F-Measure: 99.50%.	CEAS2008 (28590 spam, 11410 ham) CSDMC 2010 (1378 spam, 2949 ham)

Chapter 10

Conclusion

This thesis presented a multi-dimensional novel approach that is simple yet effective in detection and classification of phishing emails. It has shown how the unique characteristics of email headers can be exploited with n-gram analysis to produce features that can distinguish between phishing and legitimate emails. The approach in this thesis studies the performance of different classifiers on different order of n-gram features, some binary and some continuous features from several datasets. The results obtained are promising. The different systems created prove that the email is an enormous source of information that could be used for phishing detection. The header itself can provide enough data for successful classification.

Using information gain, error analysis and security analysis the results are studied in depth. Most useful features are recognized, causes of misclassification are investigated and weaknesses of each method are discussed. This thesis thus provides

a consolidated account of various email header based techniques for phishing detection. Further collaboration with Text analysis and Link analysis could result in a very powerful and useful system.

Bibliography

[1] S. Abu-Nimeh, D. Nappa, X. Wang, and S. Nair. A comparison of machine learning techniques for phishing detection. In *Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit*, pages 60–69. ACM, 2007.

[2] O. Al-Jarrah, I. Khater, and B. Al-Duwairi. Identifying potentially useful email header features for email spam filtering. In *The Sixth International Conference on Digital Society (ICDS)*, 2012.

[3] ApacheSpamAssassin. Spamassassin public mail corpus. `https://spamassassin.apache.org/publiccorpus/`, 2006.

[4] APWG. Phishing activity trends report, 4th quarter, 2014. `https://docs.apwg.org/reports/apwg_trends_report_q4_2014.pdf`, 2014.

[5] R. B. Basnet and A. H. Sung. Classifying phishing emails using confidence-weighted linear classifiers. In *International Conference on Information Security and Artificial Intelligence (ISAI)*, pages 108–112, 2010.

[6] N. M. Bianchi. The return of the open relays. `http://www.spamhaus.org/news/article/706/the-return-of-the-open-relays`, 2013.

[7] L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[8] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[9] W. B. Cavnar, J. M. Trenkle, et al. N-gram-based text categorization. *Ann Arbor MI*, 48113(2):161–175, 1994.

[10] M. Chandrasekaran, K. Narayanan, and S. Upadhyaya. Phishing email detection based on structural properties. In *NYS Cyber Security Conference*, pages 1–7, 2006.

[11] T.-C. Chen, T. Stepan, S. Dick, and J. Miller. An anti-phishing system employing diffused information. *ACM Transactions on Information and System Security (TISSEC)*, 16(4):16, 2014.

[12] B. Costales, G. Janse, C. Abmann, and G. N. Shapiro. Sendmail (4th ed.). In *Sendmail (4th ed.)*. O'Reilly, 2007.

[13] K. Crammer. Confidence weighted learning library. `http://webee.technion.ac.il/people/koby/code-index.html`, 2009.

[14] CSMiningGroup. Spam email datasets, csdmc2010 spam corpus. `http://csmining.org/index.php/spam-email-datasets-.html`, 2010.

[15] C. Fellbaum. *WordNet*. Wiley Online Library, 1998.

[16] I. Fette, N. Sadeh, and A. Tomasic. Learning to detect phishing emails. In *Proceedings of the 16th international conference on World Wide Web*, pages 649–656. ACM, 2007.

[17] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *Thirteenth International Conference on Machine Learning*, pages 148–156, San Francisco, 1996. Morgan Kaufmann.

[18] A. Genkin, D. D. Lewis, and D. Madigan. Large-scale bayesian logistic regression for text categorization. *Technometrics*, 49(3):291–304, 2007.

[19] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[20] I. R. A. Hamid and J. Abawajy. Hybrid feature selection for phishing email detection. In *Algorithms and Architectures for Parallel Processing*, pages 266–275. Springer, 2011.

[21] Y. Hu, C. Guo, E. Ngai, M. Liu, and S. Chen. A scalable intelligent non-content-based spam-filtering framework. *Expert Systems with Applications*, 37(12):8557–8565, 2010.

[22] C.-Y. Huang, S.-P. Ma, W.-L. Yeh, C.-Y. Lin, and C.-T. Liu. Mitigate web phishing using site signatures. In *TENCON 2010-2010 IEEE Region 10 Conference*, pages 803–808. IEEE, 2010.

[23] D. Irani, S. Webb, J. Giffin, and C. Pu. Evolutionary study of phishing. In *eCrime Researchers Summit, 2008*, pages 1–10. IEEE, 2008.

[24] G. H. John and P. Langley. Estimating continuous distributions in bayesian classifiers. In *Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345, San Mateo, 1995. Morgan Kaufmann.

[25] R. Layton, P. Watters, and R. Dazeley. Automatically determining phishing campaigns using the uscap methodology. In *eCrime Researchers Summit (eCrime), 2010*, pages 1–8. IEEE, 2010.

[26] B. Leiba, J. Ossher, V. Rajan, R. Segal, and M. N. Wegman. Smtp path analysis. In *CEAS*, 2005.

[27] G. F. Lyon. *Nmap Network Scanning: The Official Nmap Project Guide to Network Discovery and Security Scanning*. Insecure, 2009.

[28] A. Mejer and K. Crammer. Confidence in structured-prediction using confidence-weighted models. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 971–981. Association for Computational Linguistics, 2010.

[29] R. Mihalcea and A. Csomai. Senselearner: Word sense disambiguation for all words in unrestricted text. In *Proceedings of the ACL 2005 on Interactive poster and demonstration sessions*, pages 53–56. Association for Computational Linguistics, 2005.

[30] J. Nazario. The online phishing corpus. `http://monkey.org/~jose/wiki/doku.php`, 2004.

[31] J. Nazario. Nazario new phishing corpus (private dataset), 2014.

[32] S. Pasupatheeswaran. Email 'message-ids' helpful for forensic analysis? *School of Computer and Information Science, Edith Cowan University, Perth, Western Australia*, 2008.

[33] J. Platt et al. Fast training of support vector machines using sequential minimal optimization. *Advances in kernel methods-support vector learning*, 3, 1999.

[34] W. Primer. Url: http://weka. wikispaces. com.

[35] A. Qaroush, I. M. Khater, and M. Washaha. Identifying spam e-mail based-on statistical header features and sender behavior. In *Proceedings of the CUBE International Information Technology Conference*, pages 771–778. ACM, 2012.

[36] J. R. Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.

[37] P. Resnick. Internet message format. `http://www.ietf.org/rfc/rfc2822.txt`, 2001.

[38] F. Toolan and J. Carthy. Feature selection for spam and phishing detection. In *eCrime Researchers Summit (eCrime), 2010*, pages 1–12. IEEE, 2010.

[39] R. Verma and K. Dyer. On the character of phishing urls: Accurate and robust statistical learning classifiers. In *Proceedings of the 5th ACM Conference on Data and Application Security and Privacy*, pages 111–122. ACM, 2015.

[40] R. Verma and N. Hossain. Semantic feature selection for text with application to phishing email detection. In *Information Security and Cryptology–ICISC 2013*, pages 455–468. Springer, 2014.

[41] R. Verma, N. Shashidhar, and N. Hossain. Detecting phishing emails the natural language way. In *ESORICS*, pages 824–841, 2012.

[42] R. M. Verma and N. Rai. Phish-idetector: Message-id based automatic phishing detection. In *SECRYPT 2015 - Proceedings of the 12th International Conference on Security and Cryptography, Colmar, Alsace, France, 20-22 July, 2015.*, pages 427–434, 2015.

[43] C.-C. Wang. Sender and receiver addresses as cues for anti-spam filtering. *Journal of Research and Practice in Information Technology*, 36(1):3–7, 2004.