Proposed Discriminative Lexical Features for Real-time Detection of Malware Uniform Resource Locator

Article *in* Indian Journal of Science and Technology · December 2016

DOI: 10.17485/ijst/2016/v9i46/107081

4 authors, including:

Morufu Olalere
Federal University of Technology Minna
23 PUBLICATIONS **84** CITATIONS

SEE PROFILE

Ramlan Mahmod
Universiti Putra Malaysia
206 PUBLICATIONS **1,214** CITATIONS

SEE PROFILE

Azizol Abdullah
Universiti Putra Malaysia
140 PUBLICATIONS **821** CITATIONS

SEE PROFILE

Indian Journal of Science and Technology, Vol 9(46), DOI: 10.17485/ijst/2016/v9i46/107081, December 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Proposed Discriminative Lexical Features for Real-time Detection of Malware Uniform Resource Locator

M. Olalere[1], M. T. Abdullah[2]*, R. Mahmod[2] and A. Abdullah[2]

[1]Cyber Security Science Department Federal University of Technology Minna, 920102, Niger, Nigeria
[2]Information Security Research Group Faculty of Computer Science and information Technology, Universiti Putra Malaysia, 43400, Selangor, Malaysia; lerejide@futminna.edu.ng

Abstract

Objective: To identify discriminative lexical features of malware URL through manual examination, and to study prevalence of these features thereby leading to proposition of discriminative lexical feature for real-time detection of malware URL. **Methods/Statistical Analysis:** Manual examination of malware URL using existing blacklist of malware URLs and empirical analysis allowed the authors to identify discriminative lexical features and to determine whether there is consistency in the way the attackers craft malware URLs respectively. Empirical analysis was carried on both the existing blacklisted malware URLs and newly collected malware URLs. Empirical analysis revealed that there is consistency in the way malware URLs is crafted by the attackers. To evaluate performance of our proposed lexical features, two previously used machine learning models were applied on our trained dataset of malware URLS and benign URLs. The essence of using these models is to enable us compare performance of our proposed lexical features with previous studies proposed feature groups. Our comparison shows that our proposed lexical features outperform previously proposed feature groups. **Findings:** Our first step was to manually examine blacklisted malware URLs. This step led to the identification of 12 discriminative lexical features which was later reduced to 11. The second step was an empirical analysis of the identified features of existing blacklisted malware URLs and newly collected malware URLs. Empirical analysis was carried out to determine whether there was consistency in the way malware URLs are crafted by the attackers. The results of our empirical analysis revealed that there is indeed consistency in the way malware URLs are crafted by the attackers. This implies that our carefully identified lexical features are common features of malware URL. After experimentation, the evaluation results reveal that our proposed lexical features outperform previously proposed feature groups. **Applications/Improvements:** Discriminative features are required to build real-time malware URLs detection system with machine learning algorithm. The proposed lexical features are set of discriminative feature that rely on textual properties of malware URL.

Keywords: Attackers, Blacklist, Lexical Features, Malware URL, Rea-time Malware URL Detection

1. Introduction

Gone are the days when a malware infection on an enterprise network occurred only through external storage devices such as external hard disks and flash drives. With the rapid proliferation of Internet technologies, mobile devices, and web applications, attackers now use the Web as a vector for introducing malware into enterprise networks through employee's mobile devices in an environment such as Bring Your Own Device (BYOD). No wonder the Malware challenge remains the topmost challenge facing BYOD[1]. The personal mobile device is used to access a web application through the Internet either by typing a URL in the web browser or by clicking a URL link to the web application. In

*Author for correspondence

any case, URLs serve as a means of obtaining access to web applications, thus making it an exploitable tool for attackers to infect malware into the device of their victim.

However, this change in attack vector has forced many organisations to subscribe to blacklisting services of malware URLs which are provided by a range of techniques including manual submission of suspected malware URLs and honeypots. With 571 new websites available on the Internet per minute[2], the blacklist approach to detect malware URLs is no longer sufficient as many new malware URLs are not blacklisted immediately they are launched on the Internet. More so, since the blacklist is created by volunteer experts, human error in detection is unavoidable. Exact matching in blacklisting also renders it easy to be evaded[3].

To address blacklisting challenges, a real-time anomaly based detection of malware URLs is necessary. This approach relies on a machine learning detection model that detects malware URLs as soon as they are encountered, without having to visit the blacklist server. To build such a machine learning detection model, the features of malware URLs play an important role. The selection of discriminative features for any detection algorithm determines the performance of the algorithm. The need for the selection of discriminative features for a malware URL detection model motivated this study. It should be mentioned here that recent studies of other researchers have used different categories of features for the detection of malicious URL (especially phishing and spam). To the best of the knowledge of the authors, little work has been done in the area of malware URL detection or classification. A recent survey[4], concerning malicious URLs (phishing, spam and malware) detection techniques reported works of[3,5] as the only malware URL detection studies. Previous studies[3,6,7] used lexical features (textual properties) of URLs as discriminative features for malware URL detection in the case of[3] and phishing URL detection in the case of[6,7]. Similarly, our study proposed discriminative lexical features of malware URL.

2. Methodologies

In a situation where there are hundreds or thousands of features, the problem of selecting a subset of a relevant feature set for the best prediction accuracy is always a challenge for detection models. The detection model for malware URL is not left out of this challenge. To address this issue, we used two processes for proposing discriminative lexical features for malware URL detection. These processes include manual examination of URLs in an existing blacklist of malware URLs for identification of discriminative lexical features, and empirical analysis for studying the prevalence of identified features.

A malware patrol blacklist[8] was used to carry out a manual examination and empirical analysis. Malware patrol is a community of security experts that started operation in 2005 and it is a platform where anyone can submit a suspicious URL that may carry malware, viruses, or Trojans, or ransomware. When a URL is submitted, it is verified by security experts before it is added to the blacklist. The blacklist is updated every 1 hour for subscribers with a monthly payment subscription and every 48 or 72 hours for subscribers with a free subscription. Apart from the fact that the malware patrol blacklist was used by previous studies[7,9], the hourly update is also a factor we considered before choosing the malware patrol blacklist as a source of malware URL data for our study. To evaluate our proposed lexical features, experimentation was carried out. Experimentation involved training dataset of malware URLs and benign URLs. Previously Used machine learning models were used to evaluate our proposed lexical features. Finally, our evaluation results were compared with previous studies.

2.1 Manual Examination of Malware URLS Blacklist

To carry out the manual examination, we downloaded the malware URL blacklist from malware patrol website on the 4th August 2015. On this day, a total of 62015 malware URLs were available on the blacklist. The URLs on the blacklist were manually examined in order to identify discriminative lexical features that make the blacklist URLs different from benign URLs. The discriminative lexical features were identified from three main components (protocol, hostname, and path) of a URL as shown in Figure 1.

Based on these components, the feature set is grouped into three groups. Each group comprises of two or more features. The groups are URL to Path features, hostname features, and path features. It is important to note that the technicality behind the lexical struc-

ture of malware URL is beyond the scope of this study. Hence, the reason(s) behind the way malware URLs are crafted is/are not discussed in this study. Based on the feature set groups, the feature set identified during manual examination of the blacklisted malware URLs are presented below.

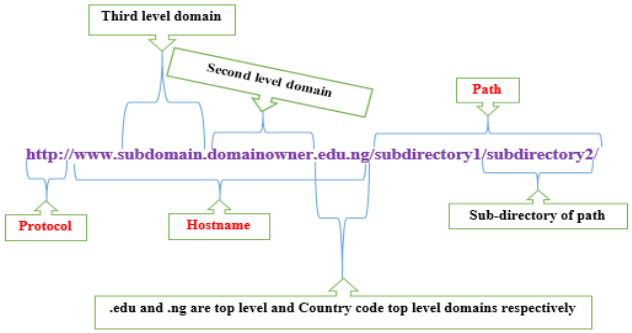

Figure 1. Components of URLs considered for feature set identification.

2.1.1 URL to Path Features Group

Two features were identified from this feature set group. These features include the following:

i. Length of URL from protocol to the path end

When we examined the URLs on the blacklist, we observed that some of the URLs have long character strings from the protocol to the end of path. Some URLs on the blacklist have as long as 250 characters. The following URL is an example of URL with long characters from the malware URL blacklist.

dde.integration.storage.conduit-services.com/39/233/ct2331539/cbdebcb46b4149109bd1ed6efbe14178/downloads/prod/dde1.3.8.4_perion.131024.04/13-11-05-21.50.02.936/

ii. Length of URL from protocol to the path end

Our manual examination revealed that many URLs have IP addresses in their hostname (the hostname is either replaced by the IP or the IP is added to hostname), path, and in some cases, both. This implies that the occurrence of the IP address in any part of the URL is a strong indication that the URL is a malware URL.

120.198.196.101/sanlixop/sanlix_data/sample/unknown/2013-03/2013-03-14/106714/

2.1.2 Hostname Features Group

Our manual examination of the blacklisted malware URLs revealed that the hostname of the malware URL is crafted in a form that is different from the hostname of benign URLs. Consequently, five discriminative lexical features were identified. These features are described below.

i. Length of Hostname

During manual examination, it was observed that many URLs have long character strings which make them different from benign URLs. Example of this type of URL from the malware URL blacklist is given below.

dde.de.resource-efiles-drive.com/29/773/ct7739229/4caee31a80f04d0a83e40d536dba48eb/Downloads/Prod/SmallStub1.3.9.0.140504.01/15-01-25-09.49.18.728/

ii. The Presence of www

Manual examination of malware URL blacklist revealed that many URLs from the blacklist do not have www. Very few URLs on the malware URL blacklist have www. All the examples of URL given above have no www. The following URL is another example of URL from the malware URL blacklist that has no www.

download2.77169.com/soft/hacrktools/attack/200906/

iii. The Presence of a Third Level Domain (TLD)

Manual examination of the malware URL blacklist revealed that many URLs on the blacklist have TLD. Example of this type of URL from the malware URL blacklist is as follows:

dl-2.one2up.com/one two/content/2014/6/12/

iv. The Presence of a Decimal Number in the Second Level Domain (SLD)

Many URLs on the malware URL black list have decimal number in their SLD. It was observed that some URLs SLDs have combination of decimal number(s) and alphabet(s). While some URLs on the malware URL blacklist have only decimal number(s) as their SLDs. Example of URL in this category is given below.

download5.77169.com/soft/other/2006/200612/

v. The Presence of a Decimal Number in the TLD

During manual examination of the malware URL blacklist, it was observed that many URLs on the blacklist have decimal numbers in their TLDs. Some of the URLs on the malware URL blacklist have only decimal number(s) as their TLDs. While some of the URLs have combination of decimal number(s) and alphabet(s) as their TLDs. URL

below is an example of this category of malware URL from the blacklist.
56ffec5e.dl-one2up.com/one two/content/2015/9/27/

2.1.3 Path Features Group

The path features group represents features identified from the path of the URL. We identified five features from the URL path. These features are described below.

i. Length of the path
The length of the path of the malware URL was observed to be long in most of the blacklisted URLs. Example of this type of URL from the malware URL blacklist is given below.
s.ddirectdownload-about.com/82/288/ct2888182/67b7b53e3fc449c8a73307c88c60bb39/Downloads/Prod/DDE1.4.0.5.150121.02/15-02-17-18.05.10.828/

ii. Number of Subdirectories in the Path
When the malware URL blacklist was examined, it was observed that many of the URLs on the blacklist have two or more subdirectories in their paths. The URL below is an example of this type of URL from the malware URL blacklist. The URL has 8 subdirectories.
s.ddirectdownload-about.com/95/242/ct2427695/ea7f8d9e06d64be6b9730677d138730f/downloads/prod/dde1.4.0.5.150121.02/15-03-07-05.40.59.238/

iii. Length of Longest Subdirectory
During manual examination, it was observed that some of the URLs on the malware URL blacklist have one or more of their subdirectories very long. Below is an example of malware URL with the longest length of its subdirectory equal to 32.
218.207.102.106/1Q2W3E4R5T6Y7U8I9O0P1Z2X3C4V5B/dlsw.baidu.com/sw-search-sp/2015_05_08_20/bind1/36561/

iv. The Presence of a Date in the Path
Many URLs on the malware URL blacklist have a date in their path. It was observed that presence of a date in the path takes different formats. Some of the URLs on the malware URL blacklist have full date format (with month, day and year), while some have only year. Example of URL with dates is given below.
60.10.0.246/1103esv2013/files/322500000016514D/dlsw.baidu.com/sw-search-sp/2015_05_08_22/bind1/11006/

v. The Presence of Hexadecimal String in the Path
The last feature identified under this group is whether there is a hexadecimal character string in the path or otherwise. We observed that many URLs on the malware URL blacklist have a hexa-decimal character string. The URL below is an example of URL with hexadecimal string in the path.
cdn1.mydown.yesky.com/55a6673a/df3b2fe-23a66e96894a7ad6e3f5ddbd3/soft/200807/

2.2 Empirical Analysis

Some of the identified features are categorical (present or not present) while others are not. These categorical features include the presence of an IP, presence of www, presence of a date, whether the hostname has a TLD or otherwise, presence of a decimal number in a SLD, presence of a decimal number in the TLD, and whether a hexadecimal character string is present in the path or not. To study the prevalence of these features, we carried out an empirical analysis of 62103 malware URLs on the blacklist and on the newly collected (as the blacklist is updated) malware URLs. The purpose of this empirical analysis was to determine the level of consistency in the way attackers craft malware URLs. Details of the empirical analysis are described in the following subsections.

2.2.1 Analysis of 62013 URLs

Under this analysis, we extracted the total number of URLs having each of the categorical features. Then, the percentage of each feature appearance in the 62103 malware URLs was computed. Table 1 shows the result of the percentage appearance of each of the categorical features in the 62103 malware URLs.

Table 1. Percentage of each of the categorical features in 62103 malware URLs

Total URL	62103		
No.	**Features**	**No. of URL**	**% in Total URL**
1	Presence of IP address	11422	**18.39**
2	Presence of www	57296	**92.26**
3	Presence of a date in the path	27388	**44.10**
4	Presence of TLD	49815	**80.21**
5	Presence of a decimal number in the SLD	17233	**27.75**
6	Presence of a decimal number in the TLD	19218	**30.95**
7	Presence of hexadecimal in path	7988	**12.86**

2.2.2 *Analysis of Newly Collected 18015 URLs*

To study the prevalence pattern in which malware URL was crafted, we collected newly added malware URLs from [8]. This collection took place from 5th August 2015 to 13th October 2015 and resulted in a total of 18015 malware URLs in 30 rounds. Table 2 summarises the details of how the URLs were collected. While in all the 30 rounds, Table 3, shows the percentage of the URLs with IP address, without www, with a date, whit a TLD, with a decimal number in the SLD, whit a decimal number in the TLD and with hexadecimal character string in the path. Meanwhile, Table 4 shows the result of the percentage appearance of each of the categorical features in the 18015 malware URLs.

Table 2. Details of how URLs were collected

Collection round	Date interval	No. of days	No. of URL
Round1	05-07/08/2015	3	**205**
Round2	08-09/08/2015	2	**149**
Round3	10-11/08/2015	2	**184**
Round4	12-14/08/2015	3	**177**
Round5	15-16/08/2015	2	**100**
Round6	17-18/08/2015	2	**47**
Round7	19-21/08/2015	3	**127**
Round8	22-23/08/2015	2	**1330**
Round9	24-25/08/2015	2	**978**
Round10	26-28/08/2015	3	**1783**
Round11	29-30/08/2015	2	**1329**
Round12	31-01/09/2015	2	**1400**
Round13	02-04/09/2015	3	**925**
Round14	05-06/09/2015	2	**457**
Round15	07-08/09/2015	2	**222**
Round16	09-11/09/2015	3	**464**
Round17	12-13/09/2015	2	**1451**
Round18	14-15/09/2015	2	**529**
Round19	16-18/09/2015	3	**1649**
Round20	19-20/09/2015	2	**329**
Round21	21-22/09/2015	2	**301**
Round22	23-25/09/2015	3	**583**
Round23	26-27/09/2015	2	**351**
Round24	28-29/09/2015	2	**368**
Round25	30-02/10/2015	3	**1018**
Round26	03-04/10/2015	2	**594**
Round27	05-06/10/2015	2	**114**
Round28	07-09/10/2015	3	**94**
Round29	10-11/10/2015	2	**71**
Round30	12-13/10/2015	2	**686**
TOTAL		**70**	**18015**

Table 3. Percentage of each of the categorical features in all the 30 rounds

Round	Total URL collected per round	% of presence of IP	% of URLs without www	% of URLs with date	% of URLs with TLD	% of URL with a decimal No. in SLD	% of URLs with a decimal No. in TLD	% of presence of URLs with hexadecimal in path
Round1	205	6.83	95.12	25.37	88.29	11.22	10.24	34.63
Round2	149	16.11	96.64	16.78	86.58	10.74	14.09	38.26
Round3	184	2.17	65.76	19.02	52.17	11.41	17.39	13.04
Round4	177	14.69	80.23	23.73	68.93	19.77	19.21	12.43
Round5	100	11.00	83.00	33.00	74.00	26.00	35.00	17.00
Round6	47	44.68	82.98	31.91	76.60	14.89	25.53	29.79
Round7	127	4.72	96.06	21.26	89.76	21.26	44.09	19.69
Round8	1330	8.65	96.77	11.80	93.38	21.28	54.74	49.77
Round9	978	3.17	97.75	13.80	92.84	21.98	55.42	62.58
Round10	1783	6.17	96.52	42.12	92.71	19.63	25.29	35.73
Round11	1329	11.29	103.16	37.40	94.73	17.91	25.66	39.95
Round12	1400	7.21	97.93	32.21	95.79	18.64	25.43	51.64

Round13	925	1.51	97.95	13.51	92.22	21.73	34.59	64.76
Round14	457	1.09	98.25	18.60	94.97	10.28	31.07	59.52
Round15	222	10.36	97.75	35.59	96.40	14.86	20.27	45.50
Round16	464	3.02	98.28	17.24	97.20	20.47	23.92	55.82
Round17	1451	3.03	96.76	9.30	91.18	17.37	23.85	26.12
Round18	529	5.67	95.84	20.98	91.87	37.24	20.42	39.89
Round19	1649	7.28	97.82	25.47	94.12	20.92	26.32	30.14
Round20	329	15.20	95.74	13.37	86.93	27.96	39.82	12.16
Round21	301	6.64	96.68	21.26	82.72	29.57	38.21	11.63
Round22	583	7.03	93.65	8.06	80.27	33.79	18.35	19.90
Round23	351	3.13	91.17	10.26	72.93	9.40	19.94	13.11
Round24	368	2.72	93.48	17.12	79.62	10.60	14.67	13.86
Round25	1018	7.86	97.15	10.71	92.83	22.89	54.72	10.31
Round26	594	14.98	97.47	17.00	93.43	24.75	23.23	10.61
Round27	114	13.16	90.35	14.91	75.44	21.93	22.81	16.67
Round28	94	10.64	86.17	11.70	74.47	13.83	38.30	13.83
Round29	71	7.04	92.96	38.03	80.28	19.72	33.80	15.49
Round30	686	15.16	92.13	17.64	79.74	9.48	28.43	8.89
Total	18015	7.21	**96.42**	**21.62**	**90.37**	**20.09**	**31.02**	**34.82**

Table 4. Percentage of each of the categorical features in 18015 malware URLs

Total URL	**18015**		
No.	**Features**	**No. of URL**	**% in Total URL**
1	Presence of IP	1298	**7.21**
2	Presence of www	17370	**96.42**
3	Presence of a date in the path	3895	**21.62**
4	Presence of TLD	16280	**90.37**
5	Presence of a decimal number in the SLD	3619	**20.09**
6	Presence of a decimal number in the TLD	5588	**31.02**
7	Presence of hexadecimal in the path	6273	**34.82**

3. Summary of Empirical Analysis

Figure 2 shows comparison of percentages of each of the categorical features in the 62103 and 18015 URLs. The percentage of the presence of decimal numbers in the TLD in the 62103 URLs was the same as the percentage of the presence of decimal numbers in the TLD in the 18015 URLs. The presence of www, presence of TLD, and presence of hexadecimal numbers in the path have almost the same percentage in both cases. Also, the percentages of the presence of an IP, presence of a date, and presence of decimal numbers in the SLD were slightly higher in the 62103 URLs than in the 18015 URLs. The implication of this is that the attackers tend to use to the same pattern of crafting malware URLs.

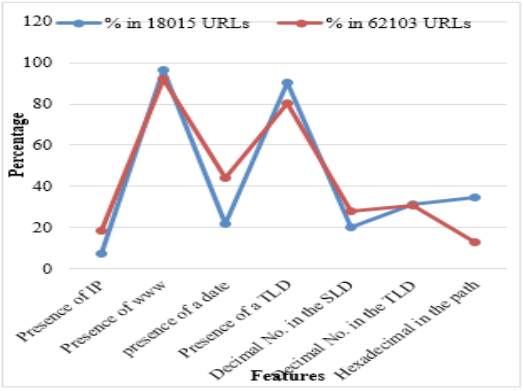

Figure 2. Comparison of percentage of each of the categorical feature in both the 62103 URLs and the 18015 URLs.

However, Figure 2 shows that more than 80 % of the 62103 and 90% of the 18015 URLs contain the TLD. This implies that many malware URLs are crafted to include the TLD. Our analysis revealed that many URLs with a decimal number in the SLD also have a decimal number in the TLD. The SLD and TLD belong to the same part (hostname) of the URL. We therefore combined the presence of a decimal number in the SLD and TLD to form a single feature. We refer to this feature as the presence of a decimal number in the hostname. Table 5 shows a summary of all features with their value type.

Table 5. Summary of the proposed features with their value type

Feature groups	Features	Value type
URL to path	Length of URL to the path end	Integer
	Presence of IP address	Binary
Hostname	Length of the hostname	Integer
	Presence of www	Binary
	Presence of a TLD	Binary
	Presence of a decimal number in the hostname	Binary
Path	Length of the path	Integer
	Number of Subdirectory in the path	Integer
	Length of longest subdirectory in the path	Integer
	Presence of a date in the path	Binary
	Presence of Hexadecimal in the path	Binary

4. Experimentation

For the purpose of evaluation and comparison of effectiveness of our proposed discriminative lexical features with previous studies, we applied two different machine leaning algorithms proposed by[3,10] on our trained dataset of malware and benign URLs. In[3], Support Vector Machine (SVM) was used to evaluate proposed discriminative lexical features and some other feature groups including Link Popularity Features (LPOP), Webpage Content Features (CONT), Domain Name System Features (DNS), DNS Fluxiness Features (DNSF), and Network Features (NET). SVM finds the hyperplane that has the largest distance to the nearest training data points of any class, called the functional margin[3]. Recently, the study by[10] proposed Naïve Bayes (NB) for detection of various malicious URLs the use of some of the discriminative features proposed by[3]. NB is a simple probabilistic classifier that is based on applying Bayes theorem from Bayesian statistics with naïve independence assumptions[10]. It is important to note that the use of these two machine learning algorithms is to enable us compare our proposed discriminative lexical features with the previously proposed other feature types.

4.1 Data Collection and Feature Extraction

The URLs collected for 70 days from[8] constituted malware URLs used for the training dataset. Benign URLs were collected from the dmoz open directory project[11]. This open directory project has become a popular source of benign URLs for malicious URL classification. The dmoz open directory project is a comprehensive open directory of the web which is manually edited by volunteer editors. The open directory project contains many categories of URLs from different topics. Hence, URLs are randomly obtained from all the URL categories in the directory. A random collection of different categories of URLs from different topics gives a true representation of a real life scenario. The link klipper web crawling tool[12] was used to obtain benign URLs from the dmoz open directory project from 14th September 2015 to 28th September 2015. Percentage of malware and benign URLs present in the dataset is presented in Table 6. Features of both benign and malware URLs on our dataset were extracted based on the feature value types described in Table 5. After features extraction, WEKA data mining tool was used to run our experiment.

Table 6. Summary of the proposed features with their value type

URL types	Number of URLs	% in total URLs
Malware	18015	43.31
Benign	23582	56.69
Total URLs	**41597**	

5. Performance Evaluation and Comparison with Previous Studies

In order to compare the performance of the proposed discriminative lexical features of this study with the

proposed/used features of the above previous studies, this study proposed features were used to train the NB and SVM. The same experimental procedures that were used by[3] were used during the authors experiment. According to[3] "Two-fold cross validation was performed to evaluate our method: the URLs in each data set were randomly split into two groups of equal size: one group was selected as the training set while the other was used as the testing set" (p. 6). Accuracy and True Positive Rate were used as evaluation parameters in[3]. For the sake of comparison, the same evaluation parameters were used in the present study. Table 7 presents the results of our evaluations when we applied the same experimental procedures in[3].

Table 7. Evaluation results of our experiment

Evaluation parameters	SVM	NB
Accuracy	96.43%	95.23%
TPR	95.58%	86.24%

5.1 Comparison of the Proposed Features with the Proposed Lexical Features of Previous Studies

As mentioned in section 4, the study by[3] proposed a discriminative lexical feature which was evaluated with SVM. Meanwhile, the study by[10] evaluated some of the discriminative lexical features proposed by[3] with NB. In the same way, we compared our results with the results of[3,10]. Table 8 shows a comparison of the performance of our proposed discriminative lexical features with the proposed/Used features of previous studies[3,10]. It can be seen from Table 8 that this study proposed discriminative lexical features performed best in terms of accuracy and TPR.

5.2 Comparison of Performance of the Proposed Lexical Features with Other Proposed Feature Groups of Previous Study

In section 4, it was mentioned that apart from the discriminative lexical features proposed by the study of[3], the authors also proposed some other feature groups. A comparison of the performance of our proposed lexical features with other feature groups proposed by[3] is presented in Table 9. As it can be seen from Table 9, this study proposed lexical features outperform all the five proposed features by3.

Over recent years, the detection and classification of malicious URLs (phishing, spam and malware) has attracted the attention of many researchers. Attention has been mostly given to phishing and spam URL classification and detection with little work on malware URL. Different feature sets for building classifiers and detectors have been proposed. For instance, However, in the field of detection or classification of malicious ULRs (spam, phishing and malware), previous studies have used different groups of feature sets. For instance,[13] in their phishing URL detection study used a different set of feature groups, which included lexical URL features, IP address properties, WHOIS properties, domain name properties, blacklist membership, geographic properties, and connection speed. The study by[3] on the detection of malicious URLs of all the popular attack types (spam, phishing, and malware) used 5 different feature groups. A study by[5] on the detection of malicious webpages pertinent to drive by-download, phishing, injection, and malware distribution attacks extracted features such as URL features,

Table 8. Comparison of this study proposed lexical features with the previous studies lexical features

Evaluation parameters	Studies					
	This study		Sayamber and Dixit (2014)		Choi et al. (2011)	
	SVM	NB	SVM	NB	SVM	NB
Accuracy	**96.43%**	95.23%	About 74.00%	About 77.00%	70.30%	Not available
TPR	**95.58%**	86.24%	Not available	Not available	74.50%	Not available

Table 9. Comparison of this study proposed lexical features with other proposed feature groups by previous study

Evaluation parameters	APFs	LPOP	CONT	DNS	DNSF	NET
Accuracy	**96.43%**	96.20%	86.20%	78.60%	68.10%	73.30%
TPR	**95.58%**	93.20%	88.40%	75.10%	74.20%	78.20%

page-source features and social-reputation features. In another phishing detection study by[14], feature groups which were based on lexical, keyword, search engine, reputation, and content were used.

Similar to our study, some previous studies on the detection or classification of malicious URLs have proved that only lexical features of any type of malicious URLs are sufficient for detection or classification[6,7,15]. In[7], it was proved that only lexical features are sufficient for detection of malicious URLs or classification (if used properly). The study by[6] used a set of lexical features for their phishing detection. Also, lexical features of URLs were used for detection of malicious domains study by[15].

The study by[3] was most closely related to our own study. Though other features were used, they evaluated the detection of all the three malicious URLs using only lexical features. Similar to our work, they used 10 lexical features to build a detection algorithm for all the malicious URLs. Also, SVM was used in their work. They claimed that their lexical features were effective in detecting phishing URLs, but did a poor job to detect spam and malware URLs. Their reason for this was that spam and malware URLs do not show very different textual patterns as compared to benign URLs. Our analysis disagrees with their stated reason. We discovered that malware URLs show very different textual patterns as compared to benign URLs.

In their study[3], the malware URLs were collected from DNS-BH[16], which was a project designed to create and maintain a list of URLs that were known to be used to propagate malware. We decided not to use[16] as a source of our malware URL because of the following: (1) when we visited the DNS-BH site, we noticed that many of the malware URL blacklisted were those with only a hostname and (2) the DNS-BH blacklist was not updated regularly, unlike reference [8]. Our first reason for not using[16] was that it might have caused poor detection accuracy for malware URL as reported by[3]. Some of the lexical features used by[3] were from the paths of URLs which were not available in[16] blacklist. Again, features such as the presence of a brand name (commonly found in phishing URL's but not in malware URL's) was among the feature set used in their work for malware URL detection. Using this type of feature gives room for biasness in the training dataset, thereby, causing poor detection accuracy.

Another closely related study to our study was the study by[10]. The study of[10] proposed a model that was based on Naïve Bayes (NB) for detection of various malicious URLs. The authors used some of the discriminative lexical features proposed by[3] and compared the performance accuracy of the NB based model with SVM based model proposed by[3]. The authors claimed that NB based model performed better than SVM model with some lexical features from the proposed lexical features by[3]. The present study has also evaluated performance of the proposed lexical feature with NB and SVM models.

6. Conclusion

In this paper, 11 novel discriminative lexical features of malware URL's are proposed. The proposed discriminative lexical features can be used to train any machine learning algorithm for real-time detection of malware URLs. Our first step was to manually examine blacklisted malware URLs. This step led to the identification of 12 discriminative lexical features which was later reduced to 11. The second step was an empirical analysis of the identified features of existing blacklisted malware URLs and newly collected malware URLs. Empirical analysis was carried out to determine whether there was consistency in the way malware URLs are crafted by the attackers. The results of our empirical analysis revealed that there is indeed consistency in the way malware URLs are crafted by the attackers. This implies that our carefully identified lexical features are common features of malware URL.

In order to evaluate and compare performance of our proposed lexical features with previous studies proposed feature groups, we ran experiment with our trained dataset of malware URLs and benign URLs using NB and SVM. The evaluation result show that our proposed lexical features outperform previous study proposed lexical features and other feature groups in terms of accuracy and TPR. Meanwhile, future study could possibly examine other machine learning algorithm on the proposed lexical features in this study.

7. References

1. Olalere M, Abdullah MT, Ramlan M, Abdullah A. A review on bring your own device on security issues. Sage Open. 2015; 05(02). p.1–11.
2. Ever wondered how many websites are created every minut? http://www.designbyconet.com/2014/06/ever-wondered-how-many-websites-are-created-every-minute/. Date accessed: 11/06/2014.

3. Choi HS, Zhu BB, Lee H. Detecting malicious web links and identifying their attack types. Proceedings of the 2nd USENIX Conference on Web Application Development; 2011 USENIX Association Berkeley, CA, USA. ACM Digital Library; 2011. p. 1–11.
4. Patil DR, Patil JB. Survey on malicious web pages detection techniques. International Journal of U- and E- service, Science and Technology. 2015; 08(5):195–206.
5. Eshete B, Villafiorita A, Weldemariam K. BINSPECT: Holistic analysis and detection of malicious web pages. Proceedings of 8th International ICST Conference, SecureComm 2012; 2012 Sep. 3–5; Padua, Italy. Berlin: Springer; 2013. p. 149–166.
6. Blum A, Wardman B, Solorio T, Warner G. Lexical feature based phishing URL detection using online learning. Proceedings of the 3rd ACM workshop on Artificial Intelligence and Security; 2010 October 04 – 08; Chicago, Illinois, USA. ACM; 2010. p. 54–60.
7. Le A, Markopoulou A, Faloutsos M. PhishDef: URL names say it all. Proceedings of IEEE INFOCOM, 2011; 2011 April 10-15; Shanghai, China. IEEE; 2011. p. 191–195.
8. Malwarepatrol. http://www.malwarepatrol.net. 2005.
9. Kalafut AJ, Shue CA, Gupta M. Malicious hubs: detecting abnormally malicious autonomous systems. Proceedings of IEEE INFOCOM, 2010 Conference. 2010 March 14-19; San Diego, CA, USA. IEEE; 2010. p. 1–5.
10. Sayamber AB, Dixit AM. Malicious URL detection and identification. International Journal of Computer Applications. 2014 August; 99(17). p.17–23.
11. dmoz. Open directory project. http://www.dmoz.org.1998-2016.
12. Link klipper. https://chrome.google.com/webstore/category/apps?hl=en.
13. Ma J, Saul LK, Savage S, Voelker GM. Beyond blacklists: Learning to detect malicious web sites from suspicious urls. Proceedings of the 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining; June-July 2009; Paris, France, ACM; 2009. p. 1245–1235.
14. Basnet RB, Sung AH. Learning to detect phishing webpages. Journal of Internet Services and Information Security (JISIS). 2014; (4)3. p. 21–39.
15. Wang W, Shirley, KE. Breaking bad: detection malicious domain using word segmentation. Proceedings of the 9th Workshop on Web 2.0 Security and Privacy (W2SP) 2015; May 21, 2015; San Jose, CA, USA. IEEE; 2015. p. 1–7.
16. DNS-BH. Malware prevention through domain blocking. http://www.malwaredomains.com. 2016.