Investigando o uso de Características na Detecção de URLs Maliciosas

Maria Azevedo Bezzera[1], **Eduardo Feitosa**[1]

[1]IComp/UFAM, Manaus, Brasil

azevedo_maria@msn.com, efeitosa@icomp.ufam.edu.br

Abstract. *Malicious URLs became a powerful channel for criminal activities on the Internet. Current solutions for URL verification present high accuracy rates with well-adjusted results, an important question can and should be done: Is it really possible to obtain 100% of accuracy in these solutions? This paper presents an investigation of features, bases and URLs formats, aiming to show that the results of URLs validation and verification are quite dependent on certain aspects and factors. By extracting features (lexical, DNS and others) directly from the URL, machine learning algorithms were employed to question their influence in the process of URLs validation and verification. Thus, four (4) cases were prepared and the evaluation shows that it is possible to disagree with the results of several studies from the literature.*

Resumo. *URLs maliciosas tornaram-se um canal poderoso para atividades criminosas na Internet. Embora as atuais soluções para verificação de URLs apresentem altas taxas de precisão, com resultados bem ajustados, um questionamento pode e deve ser feito: será que realmente é possível ou factível se obter percentuais beirando 100% de precisão nessas soluções? Neste sentido, este artigo conduz uma investigação de características, bases e formatos de URLs, visando mostrar que os resultados de validação e verificação de URLs são bastante dependentes de certos aspectos e fatores. Através da extração de características (léxicas, DNS e outras), diretamente da URL, algoritmos de aprendizagem foram empregados para questionar a influência dessas características no processo de validação e verificação de URLs. Assim, quatro (4) hipóteses foram elaboradas e a avaliação mostra que é possível discordar dos resultados de vários trabalhos já existentes na literatura.*

1. Introdução e Motivação

O sucesso de atividades maliciosas na Internet tem como ponto de partida a existência de usuários desavisados e despreparados que visitam sites desconhecidos, acessam e-mails não solicitados, ativam links e/ou fazem o download de programas de forma inadvertida. Em comum, todas essas formas de atividades maliciosas utilizam URLs (*Uniform Resource Locator*) como canal para contaminação. Dados da Kaspersky Lab [Maslennikov and Namestnikov 2012] mostram que, em 2012, 87,36% dos ataques empregaram URLs. Já o Grupo de Trabalho Anti-Phishing (APWG) relata que pelo menos 0,01% das URLs acessadas a cada dia são maliciosas [Aaron and Rasmussen 2014], o que corresponde a milhões de URLs.

Embora existam diversas soluções que visam informar ao usuário se uma URL é ou não perigosa, especialmente fazendo uso de listas negras (*blacklists*), as abordagens baseadas em aprendizagem de máquina vem ganhando espaço [Ma et al. 2009, Eshete et al. 2013, Choi et al. 2011]. A ideia geral é extrair rapidamente características da URL (caracteres, dados do host, entre outros), codificá-los e então utilizá-los para treinar uma máquina de aprendizagem a fim de construir classificadores capazes de distinguir URLs perigosas.

Dois aspectos que chamam atenção nessas soluções são a quantidade de características utilizadas para classificar uma URL e a eficácia obtida. Analisando a literatura é possível enumerar mais de 75 características extraíveis de uma URL que podem ser aplicadas na sua classificação. Tal fato gera alguns questionamentos:

1. Existem informações valiosas em todas essas características?
2. Todas essas características são necessárias e/ou são realmente utilizadas na classificação de URLs?
3. Todas essas características têm potencial para indicar ameaças?
4. Existe alguma influência da URL (formato, serviço a que se refere, base de onde foi extraída) sobre as características e, consequentemente, sobre a classificação?
5. É possível categorizar características de modo a permitir o uso mais adequado das mesmas no processo de classificação?

Diante deste contexto, o objetivo deste artigo é investigar a capacidade de validar e classificar URLs como benignas ou suspeitas/maliciosas. Para tanto, conjuntos de características extraídas das próprias URLs serão empregados como fontes de informação e diferentes métodos de aprendizagem de máquina serão utilizados para avaliação.

No que diz respeito a contribuições, este trabalho apresenta: (i) um esquema para agrupamento das características em URLs utilizadas na detecção de atividades suspeitas e/ou maliciosas; (ii) um conjunto de *scripts* que possibilitem a extração de características de URLs; (iii) indícios que fatores como o formato e o local de extração podem interferir consideravelmente no resultado do processo de classificação de URLs.

2. URL e suas características

URL (*Uniform Resource Locator*) é um formato universal para representar um recurso na Internet, de modo a ser facilmente lembrado pelos usuários. Definida e especificada na RFC 1738 [Lee et al. 1994], uma URL é composta por duas seções:

<esquema>:<parte-específica-do-esquema>

O **esquema** de uma URL representa a linguagem ou o protocolo utilizado para comunicação. No caso deste trabalho, o protocolo é o HTTP (*HyperText Transfer Protocol*) e as partes específicas são: domínio e caminho. O **domínio**, também chamado de máquina ou host, faz referência ao nome do domínio que hospeda o recurso pedido e pode ser representado tanto por um nome quanto pelo endereço IP do servidor. Um domínio é formado por um ou mais marcadores (camadas) que são concatenados e delimitados por pontos (".") e cuja hierarquia de leitura é definida da direita para a esquerda. Assim, um domínio tem em sua primeira camada, na sua parte mais a direita, um TLD (*Top Level Domain*) para representar seu tipo

(*.com*, *.net*, *.org*, *.edu*, entre outros) - chamado de *Generic TLD* (GTLD) - e o país de origem (*.br* para o Brasil, *.uk* para o Reino Unido, *.us* para os Estados Unidos, entre outros) - chamado de *Country Code TLD* (CCTLD). Em seguida, tem-se o segundo nível (*Second Level Domain* - SLD) que representa o nome do domínio propriamente dito. É possível existir ainda outras camadas cuja finalidade é representar especificidades do nome de domínio.

Já o **caminho** permite ao servidor conhecer o lugar onde o recurso está armazenado, ou seja, o(s) diretório(s) e o nome do recurso pedido, bem como os argumentos empregados para realização de alguma ação. O caminho é delimitado por uma barra ("/") e sua hierarquia de leitura é da esquerda para a direita. A Tabela 1 ilustra um exemplo completo de URL.

Tabela 1. Componentes de uma URL

Componente	Exemplo
URL	http://icomp.ufam.edu.br/inst/hstart.php?id=664&logon=141
Nome do domínio	icomp.ufam.edu.br
Caminho	inst/hstart.php
Sub diretório	inst
Nome do arquivo	hstart
Extensão do arquivo	php
Argumento	id=664&logon=141

2.1. Características de URLs

A revisão na literatura de validação e detecção de URLs, especialmente quando relacionadas as atividades maliciosas, mostra a existência de diversos trabalhos nesta área [Akiyama et al. 2011, Prakash et al. 2010, bin Lin 2008, Zhang et al. 2008, Ma et al. 2009, Eshete et al. 2013, Choi et al. 2011]. Em comum, todos eles utilizam características observáveis das próprias URLs, como, por exemplo, o tamanho e a quantidade de determinados caracteres para avaliar e inferir sobre a reputação de uma URL.

Como não existem esquemas formais para o agrupamento das características de uma URL, este trabalho as organizou de acordo com a forma de extração em offline e online (Figura 1). A Tabela 2 descreve em detalhes os grupos definidos.

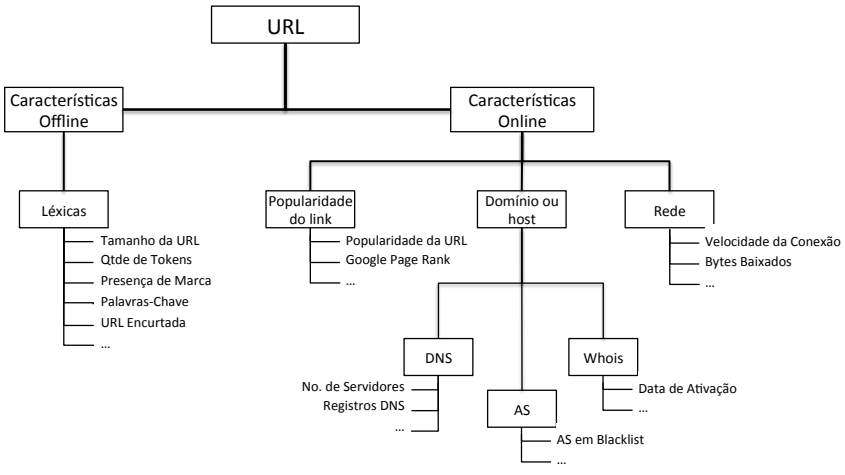

Figura 1. Agrupamento Proposto das Características

Tabela 2. Grupos de Características Definidos

Grupo	Descriçao
Popularidade do Link	Características de popularidade do link tentam estimar, através da contagem do número de ligações (links) recebidas de outras páginas Web, a importância (utilização) de uma determinada página. Em linhas gerais, tais características podem ser consideradas como uma medida da reputação de uma URL. Desta forma, enquanto sites maliciosos tendem a ter um valor pequeno de popularidade do link, sites benignos, principalmente os populares, tendem a ter um grande valor.
Relativas ao Domínio ou Host	Características relacionadas ao domínio ou host são aquelas referentes as informações exclusivas do nome de domínio ou servidor da URL. Basicamente, respondem a perguntas como "onde" está hospedado e localizado o site, "quem" o gerencia e "como" ele é administrado. Essas características são obtidas através de dados do DNS (*Domain Name Service*), do ASN (*Autonomous System Number*) e do registro (*whois*). Sua importância se deve ao fato de que sites maliciosos tendem a serem hospedados em fornecedores de serviço menos respeitáveis, em máquinas não convencionais ou em registros corrompidos.
Recursos de Rede	Esse grupo abrange aquelas características relacionadas a informações mais diversificadas, que não podem ser categorizadas nos outros grupos. Por exemplo, URLs maliciosas podem redirecionar o usuário até atingir o local da atividade ilícita. Isso pode ocorrer através de redirecionamentos dentro do código HTML ou via o encurtamento das URLs. Além dessa características, informações que precisam contabilizar respostas vindas de serviços Internet também compõem esse grupo.
Léxicas	Características léxicas são as propriedades textuais que compõe uma URL, incluindo os símbolos e marcadores, mas não incluindo o conteúdo da página. Uma vez que estão relacionadas a padrões no texto, essas características são extraídas através de tokens (símbolos como "/", ".", "," , "=" , "?", "-", "@", "&" e ou palavras-chave) da URL e empregadas em algum tipo de contabilização. De forma geral, características léxicas em URLs maliciosos tendem a "parecer diferente" aos olhos dos usuários que as vêem.

3. Trabalhos Relacionados

Esta seção realiza uma análise de alguns trabalhos relacionados à detecção de URLs maliciosas. Os trabalhos aqui apresentados representam pesquisas que utilizaram características extraídas das URLs e técnicas de aprendizagem de máquina no processo de implementação/validação. Ao final da seção, uma discussão sobre esses trabalhos é apresentada.

3.1. Trabalhos

O artigo de Ma et al. [Ma et al. 2009] descreve uma técnica para identificação de URLs suspeitas em larga escala e online, através de características léxicas e do domínio. Com base nessas características, os autores avaliaram a acurácia de quatro (4) classificadores - Naive Bayes, SVM com kernel linear (SVM-lin), SVM com kernel RBF (SMV-RBF) e regressão logística (RL) - em dados extraídos das bases DMOZ (`http://www.dmoz.org`) e Yahoo! (`https://dir.search.yahoo.com`), ambas benignas, e PhishTank (`http://www.phishtank.com`) e Spamscatter [Anderson et al. 2007] como malignas. No final, o classificador RL teve melhor desempenho que os SVMs. Infelizmente, os resultados disponibilizados não apresentam resultados separados por função/característica.

Eshete et al. [Eshete et al. 2013] aborda o projeto, a implementação e a avaliação experimental de um sistema chamado BINSPEC, que faz análise estatística e emulação para diferenciar URLs benignas de maliciosas. As características utilizadas foram divididas em URL, código fonte e reputação social, perfazendo um total de 39 elementos. De acordo com os autores, seis (6) características são novas: tamanho do caminho da URL, tamanho da consulta na URL (parte do caminho sem o diretório) e tamanho do arquivo no caminho da URL que são ligadas a URL, enquanto três (3) características de reputação social avaliam a popularidade da URL quando compartilhada no Facebook, no Twitter e no Google Plus.

No quesito avaliação, os autores utilizaram 71.919 URLs maliciosas, coletadas das blacklists do Google (`http://code.google.com/apis/`), da base PhishTank e

de uma lista de Malware (`http://www.malwareurl.com/`). Também foram usadas 414.000 URLs benignas dos sites Alexa (`http://www.alexa.com/topsites`), Yahoo! e DMOZ. Os autores empregaram sete (7) classificadores: J48, Random Tree, Random Forest, Naive Bayes, redes Bayesscianas, SVM e Regressão Logística. Como resultado, as três (3) novas características de URL melhoraram o desempenho global de 5 dos 7 classificadores (J48, Random Forest, Naive Bayes, redes Bayianas e regressão logística). Já as características de reputação social melhoraram a exatidão de classificação da Random Forest, redes Bayianas e regressão logística. Além da contribuição individual das novas características, os autores também mediram a melhoria global na precisão dos classificadores. De forma resumida, foram obtidos ganhos de precisão em 4 dos 7 classificadores, com melhorias no intervalo de 0,21% a 3,08%.

Os autores em [Fei] apresentam um método de classificação automática de XSS em páginas Web baseado em técnicas de aprendizagem de máquina supervisionadas. Embora não seja um trabalho exclusivamente focado na detecção de URLs maliciosas, e sim de conteúdo Web malicioso, apresenta uma série de características interessantes e extraíveis da URL, bem como faz uso de vários classificadores de aprendizagem de máquina. Dentre as características mais interessantes para detecção de URLs, o trabalho apresenta: (i) Código Ofuscado; (ii) Quantidade de Domínios; (ii) Caracteres Especiais Duplicados.

Em relação a avaliação e resultados, foram utilizados os classificadores Naive Bayes e SVM. Os autores montaram uma base contendo 57.207 páginas benignas da DMOZ, 158.847 páginas benignas da ClueWeb09 (`http://www.lemurproject.org`) e 15.366 páginas infectadas com código XSS da base XSSed, (`http://www.xssed.com`) referentes a ataques ocorridos de 23 de junho de 2008 a 02 de agosto de 2011. Como resultado, os autores observaram que o classificador SVM obteve os melhores valores na classificação de páginas XSS, com precisão de 98,58% para base DMOZ/XSSed e 99,89% para base ClueWeb/XSSed. Contudo, os autores comentam que o desempenho do classificador Naive Bayes foi bem próximo ao de SVM, devido ao fato de que as características propostas são aderentes ao conceito de independência condicional, ou seja, o valor de um atributo para uma classe independe dos valores dos outros atributos [Fei].

Choi et al. [Choi et al. 2011] propõem um método que utiliza a aprendizagem de máquina para detectar se uma URL é maliciosa ou não, classificando-as de acordo com tipos de ataque (*spam*, *phishing* e *malware*). Para tanto, adota um grande conjunto de características discriminativas relacionadas a padrões textuais, estruturas do link, composição do conteúdo, informações de DNS e tráfego de rede. Dentre as características léxicas abordadas no trabalho, a que trata da presença de *spam*, *phishing* e *malware* no SLD da URL é a mais inovadora.

Na avaliação, o método proposto utiliza SVM para classificar a URL como benigna ou não, e os algoritmos *multi-label* RaKEl (*Random k-labelsets*) [Tsoumakas and Vlahavas 2007] e ML-KNN (*Multi-label k-Nearest Neighbor*) [Zhang and Zhou 2007] para identificar os tipos de ataques. No que diz respeito aos resultados, foram avaliadas 72.000 URLs, onde 40.000 foram consideradas benignas e 32.000 maliciosas, com uma precisão de mais de 98% na detecção de URLs maliciosas e uma precisão de mais de 93% na identificação de tipos de ataque. Além

disso, estudou-se a eficácia de cada grupo de características discriminativas em ambos, detecção e identificação.

O trabalho de Canali et al. [Canali et al. 2011] consiste em construir um filtro, chamado Prophiler, que utiliza técnicas de análise estática para examinar rapidamente uma página Web para averiguar se ela possui ou não conteúdo malicioso. As características avaliadas pelo Prophiler são dividas em duas categorias: conteúdo da página e URL da página (o que engloba características léxicas e do host). No trabalho, os autores afirmam que além das características já empregadas em outros trabalhos, foram elaboradas mais 48, sendo 19 referentes a códigos HTML, 25 a códigos JavaScript e 4 referentes a URL.

Os algoritmos de aprendizagem de máquina utilizados foram Random Tree, Random Forest, Naive Bayes, Regressão Logística, J48 e redes Baysianas. Todos os classificadores foram testados com bases de treinamento, usando validação cruzada com 10 (dez) partições . Foi empregada uma base de dados contendo 153.115 páginas, sendo 139.321 benignas e 13.794 maliciosas, submetidas ao Wepawet (`https://wepawet.iseclab.org`) num período de 15 dias. Em média, o Prophiler produziu uma taxa de 10.4% de falsos positivos e 0.54% de falsos negativos, o que na base de validação representa descartar imediatamente 124.906 páginas benignas e poupar recursos no processo de análise.

3.2. Discussão

A discussão feita nesta seção trata dos quatro (04) classificadores (SVM, Naive Bayes, Árvore de Decisão - J48 - e KNN) empregados na confecção deste trabalho e de como eles são empregados em algum dos trabalhos relacionados apresentados.

O SVM e o Naive Bayes são os classificadores mais usados. Suas escolhas se devem ao fato de ambos conseguirem lidar bem com grandes conjuntos de dados, possuirem um processo de classificação rápido com baixa probabilidade de erros de generalização e serem bastante empregados na análise do tráfego Internet. Já a Árvore de Decisão (J48), tem a premissa de obter regras que explicam claramente o processo de aprendizagem. Por fim, o KNN tem como vantagens a simplicidade e facilidade de implementação.

4. Implementação e Protocolo Experimental

Este Seção descreve, em duas subseções, os aspectos essenciais para alcançar o objetivo de investigar a capacidade de validar/classificar URLs como benignas, suspeitas ou maliciosas. A primeira subseção descreve as características selecionadas para extração, bem como a obtenção de seus valores. A segunda subseção relata o protocolo experimental necessário (ambiente de teste, base, ajuste nos classificadores, entre outros aspectos) para, efetivamente, validar as características.

4.1. Características

Para realização deste trabalho as características selecionadas para serem extraídas das URLs, por questões de implementação, foram categorizadas em: Léxicas, DNS e especiais (características que dependem de serviços Internet ou de conexão a Internet, e não ligadas ao DNS), totalizando 56 elementos.

No que tange a implementação, *scripts* foram desenvolvidos para extração de cada uma das características, de acordo com sua funcionalidade. Para isso, a

linguagem de Perl [Raymond 1998] foi utilizada por apresentar simplicidade, portabilidade, versatilidade e capacidade de lidar com *strings*. É importante esclarecer que algumas características e suas respectivas implementações necessitam de conexão à Internet, de forma estável, para um correto funcionamento. Um bom exemplo são as três (3) características elaboradas em diferentes sub-rotinas que utilizam o módulo "Net::DNS::Resolver" do Perl para obtenção dos endereços IP e servidores de nomes.

A Tabela 3 apresenta todas as características extraídas e implementadas.

Tabela 3. Características Implementadas

Características Léxicas					
Nome	**Descrição**	**Nome**	**Descrição**	**Nome**	**Descrição**
qt_dom_ponto	Qtde de (.) no domínio	*qt_dom_hifen*	Qtde de (-) no domínio	*qt_dom_underline*	Qtde de (_) no domínio
qt_url_ponto	Qtde de (.) na URL	*qt_url_barra*	Qtde de (/) na URL	*qt_url_interrog*	Qtde de (?) na URL
qt_url_igualdade	Qtde de (=) na URL	*qt_url_hifen*	Qtde de (-) na URL	*qt_url_underline*	Qtde de (_) na URL
qt_url_arroba	Qtde de (@) na URL	*qt_url_ecomerc*	Qtde de (&) na URL	*qt_url_exclam*	Qtde de (!) na URL
qt_url_til	Qtde de () na URL	*comp_dominio*	Comprimento do domínio	*comp_url*	Comprimento da URL
qt_dir_ponto	Qtde de (.) no diretório	*qt_dir_barra*	Qtde de (/) no diretório	*qt_dir_interrog*	Qtde de (?) no diretório
qt_dir_igualdade	Qtde de (=) no diretório	*qt_dir_hifen*	Qtde de (-) no diretório	*qt_dir_underline*	Qtde de (_) no diretório
qt_dir_arroba	Qtde de (@) no diretório	*qt_dir_exclam*	Qtde de (!) no diretório	*qt_dir_til*	Qtde de (~) no diretório
qt_arq_ponto	Qtde de (.) no arquivo	*qt_arq_interrog*	Qtde de (?) no arquivo	*qt_arq_igualdade*	Qtde de (=) no arquivo
qt_arq_hifen	Qtde de (-) no arquivo	*qt_arq_underline*	Qtde de (_) no arquivo	*qt_arq_arroba*	Qtde de (@) no arquivo
qt_arq_exclam	Qtde de (!) no arquivo	*qt_arq_til*	Qtde de () no arquivo	*qt_par_ponto*	Qtde de (.) no parâmetro
qt_par_barra	Qtde de (/) no parâmetro	*qt_par_interrog*	Qtde de (?) no parâmetro	*qt_par_igualdade*	Qtde de (=) no parâmetro
qt_par_hifen	Qtde de (-) no parâmetro	*qt_par_underline*	Qtde de (_) no parâmetro	*qt_par_arroba*	Qtde de (@) no parâmetro
qt_par_ecomerc	Qtde de (&) no parâmetro	*qt_par_exclam*	Qtde de (!) no parâmetro	*qt_par_til*	Qtde de (~) no parâmetro
qt_params	Qtde de parâmetros na URL	*pres_tld_arg*	Presença de TLD no argumento da URL	*comp_diretorio*	Comprimento do diretório da URL
comp_arquivo	Comprimento do arquivo na URL	*comp_params*	Comprimento dos parâmetros da URL	—	—-
Características de DNS					
Nome	**Descrição**	**Nome**	**Descrição**	**Nome**	**Descriçã0**
ip_asssociado	No. de IPs resolvidos	*sn_associado*	No. de servidores de nome resolvidos	*data_tempo_ativo*	Tempo (em dias) de ativação do domínio
Características Especiais					
Nome	**Descrição**	**Nome**	**Descrição**	**Nome**	**Descrição**
mal_phi	Presença em listas de Phishing ou Malware	*presenca_marca*	Presença de marca	*geo_localizacao*	Localização geográfica do domínio
rank_google	Page Rank do Google	*rank_alexa*	Page Rank do Alexa	*rbl_check*	Presença do domínio em RBL (*Real-time Blackhole List*

4.2. Configurações do Ambiente de Experimentação

Para investigar a capacidade de validar URLs como benignas ou suspeitas/maliciosas foi necessária a realização de vários experimentos com os classificadores Naive Bayes, KNN, SVM e Árvore de Decisão.

Os experimentos realizados foram executados em duas máquinas. A primeira foi um notebook com sistema operacional Windows 7 64 bits, 4 GB de memória RAM, disco de 500 GB e um processador Intel Core i5, 2.3 Ghz. A segunda foi uma estação de trabalho Intel Core 7 de 3.4 Ghz, com 8 GB de memória RAM, disco de 500 GB e plataforma Linux, distribuição Ubuntu 14.04. Para a execução dos algoritmos de classificação e análise do conhecimento foi utilizado o ambiente Weka

(`http://www.cs.waikato.ac.nz/ml/weka/`), em sua versão 3.6.10, tanto para Windows quanto Linux.

Quanto aos dados, foram utilizadas três (3) bases: DMOZ, PhishTank e Shalla's Blacklist (`http://www.shallalist.de`). Esta última é uma coleção de listas de URL agrupadas em várias categorias destinadas ao uso em filtros de URL. A base DMOZ corresponde a URLs benignas e as bases PhishTank e Shalla's Blacklist correspondem a URLs maliciosas. Para o treinamento e ajuste dos parâmetros dos classificadores foram utilizadas 20.092 URLs, sendo 10.046 oriundas da base DMOZ e 10.046 da base PhishTank. Já para a etapa de teste, foram utilizadas dois conjuntos de URLs. O primeiro é composto por 20.000 URLs das bases DMOZ e PhishTank, divididas de forma igualitária. O segundo é composto por 20.000 URLs das bases DMOZ e Shalla's, também divididas de forma igualitária.

No que diz respeito a medidas de desempenho, a métrica empregada para a avaliação dos resultados foi a validação cruzada com 10 (dez) partições para os quatro classificadores, mantendo-se a mesma proporção em todos os experimentos a fim de permitir a comparação dos resultados obtidos. As medidas empregadas para a análise de desempenho foram: (i) Taxa de detecção = VP/(VP+FN); (ii) Taxa de precisão = (VP+VN) / (VP+VN+FP+FN); e (iii) Taxa de falso alarme = FP/ (FP+VN), onde VN (Verdadeiro Negativo) indica instâncias (URLs) normais classificadas corretamente; FN (Falso Negativo) indica instâncias maliciosas classificadas como normais; FP (Falso Positivo) indica instâncias normais classificadas como maliciosas; e VP (Verdadeiro Positivo) indica instâncias maliciosas classificadas corretamente.

4.2.1. Ajustes e Escolha do Melhor Classificador

Para obter o melhor resultado sobre o conjunto de dados para cada classificador, foram realizados treinamentos onde os valores dos principais parâmetros de cada classificador foram ajustados até a obtenção do valor mais adequado. Após o ajuste dos parâmetros foi possível realizar a comparação entre os resultados dos classificadores Naive Bayes, KNN, SVM e Árvore de Decisão (J.48) - (Tabela 4), a fim de determinar o classificador que melhor se ajusta ao conjunto de treinamento.

Tabela 4. Comparação entre os Classificadores

Classificador	**Naive Bayes**	**SVM**	**Árvore de Decisão**	**KNN**
Parâmetros Ajustados	-	C=50, Kernel Polinomial, Grau do Polinômio=1.0	Fator de Confiança=0,25	KNN=1
Taxa de Precisão	76,00%	91,40%	95,10%	94,90%
Taxa de Detecção	66,35%	91,43%	95,11%	94,91%
Falso Alarme	33,60%	8,60%	4,90%	5,10%

Como observado, o classificador Naive Bayes apresentou as piores taxas de precisão e de detecção e a maior taxa de falso alarme. A principal razão para isso são os tipo de dados processados. O Naive Bayes, em termos gerais, não apresenta bom desempenho para dados contínuos, lidando melhor com dados discretos. Já os outros três classificadores obtiveram altas taxas de precisão. Entretanto, essa métrica não

pode ser avaliada isoladamente, pois seu resultado expressa o percentual de exemplos classificados corretamente, independentemente da classe a qual pertence; o que pode esconder o erro de classificação da classe minoritária.

Com base nos dados obtidos, o classificador J.48 (Árvore de Decisão) foi o mais ajustado, obtendo uma taxa de 95,10% de precisão e 95,11% de taxa de detecção, além de um baixo índice de falso alarme (4,90%,), considerando o conjunto de dados fornecidos. O sucesso desse classificador, dentre outros fatores, pode estar relacionado ao fato das características apresentarem grande potencial discriminante e ao fato das estatísticas obtidas no conjunto de teste se assemelharem às estatísticas do conjunto de treinamento, pois este classificador tem acesso à probabilidade conjunta dos atributos que a árvore considera mais relevante.

Os resultados apresentados nestas análises iniciais ratificam vários trabalhos nessa área que utilizam a aprendizagem de máquina em métodos de detecção de URLs maliciosas. Além disso, corroboram com a relevância das características utilizadas, as quais foram empregadas com classificadores estáveis e de amplo uso na área de aprendizagem de máquina, o que permite inferir que tais características são aderentes e adequadas para a detecção de URLs maliciosas no contexto do conjunto de dados avaliado.

5. Hipóteses e Provas

Como já mencionado, o objetivo deste trabalho é investigar a capacidade de classificar URLs como benignas e suspeitas/maliciosas através da extração de características das URLs e sua análise em diferentes métodos de aprendizagem de máquina. Assim, a partir desse objetivo foram definidas as seguintes hipóteses que precisam de investigação:

- **H1. Existe alguma influência do formato da URL na extração das características e, consequentemente, no processo de avaliação?**
- **H2. Todas as características extraídas são realmente necessárias no processo de detecção de URLs?**
- **H3. Grupos de características permitem resultados adequados, e até melhores, no processo de detecção de URLs se comparados com características individuais**.
- **H4. A importância das características depende da base onde as URLs são coletadas**.

Esta seção apresenta as provas (experimentos e resultados) para cada uma dessas hipóteses realizadas com os dados (características) extraídos das URLs.

5.1. Provas

Formato da URL

Para avaliar a influência do formato da URL na extração das características (hipótese H1), as URLs foram divididas em domínio e caminho, onde esta última foi subdividida em diretório, arquivo e argumento. Para poder avaliar a hipótese H1, essas características são comparadas. A Tabela 5 apresenta as médias de algumas características extraídas nas URLs das duas bases maliciosas PhishTank e Shalla's Blacklist.

É fácil notar na tabela que existe uma certa diferença entre as características das bases. As relacionadas ao tamanho são maiores na base PhishTank devido ao

Tabela 5. Média de algumas Características em bases maliciosas

Características	Shalla's (Dez. 2014)	PhishTank (Nov. 2013)
comp_url	31,0367	54,1183
comp_dominio	14,3585	19,0749
comp_diretorio	6,1998	21,6629
comp_arquivo	5,3814	7,1557
comp_params	3,097	6,2263
qt_tok_dir_barra	1,7573	2,9908
sn_assoc	2,4993	1,8551
rank_alexa	129,06	16430,6377
rank_google	2,0873	0,359

fato da base conter URLs ligadas a *phishing*, enquanto a base Shalla's é mais genérica. A única excessão é a característica *sn_assoc*, o que é, na verdade, um ponto positivo, pois significa que as URLs extraídas tem, em média, 2.4 servidores de nomes respondendo pelo domínio, ou seja, é mais confiável que os domínios na base PhishTank. Já características de popularidade como *rank_alexa* e *rank_google* apresentam grandes distorções. Enquanto na base Shalla's o valor médio do *rank_alexa* é de 129, na base PhishTank esse valor é de 16430,638, o que representa uma diferença de aproximadamente de 175%. Vale lembrar que para essa característica, quanto menor o valor, mais conhecida é a URL. No caso de *rank_google*, quanto maior o valor, mais conhecida é a URL.

Desta forma, este trabalho adota a posição que *o formato da URL possui certa influência na extração de características e, consequentemente, no processo de avaliação.*

Análise das Características: Individual ou em grupos?

De forma a provar as hipóteses H3 e H4, uma série de experimentos envolvendo os quatro (4) classificadores foi realizada para avaliar a contribuição e o impacto do conjunto de características sobre o resultado final da classificação. Para tanto, foram gerados três (3) grupos formados por: (i) Características de DNS; (ii) Características Especiais; e (iii) Características Léxicas. Em termos práticos, esses grupos foram separados da seguinte forma para a validação da hipótese (Tabela 6):

Tabela 6. Grupos de Características

Conjunto	Descrição	Conjunto	Descrição
A	Composto pelas 3 características baseadas em informações obtidas do DNS	B	Composto pelas 6 características denominadas especiais
C	Composto pelas 15 características léxicas mais comuns [1]	D	Composto por 32 características léxicas variáveis (não aparecem em todas as bases de dados)
E	Composto por todas as características léxicas	A+B+C	Composto pelos grupos A, B e C, totalizando 24 características
A+B+D	Composto pelos grupos A, B e D, totalizando 41 características	Todos	Todas as 56 características

É importante enfatizar que foram utilizadas 20.000 URLs, sendo 10.000 da base DMOZ e as outras 10.000 da base PhishTank. Contudo, embora as bases sejam as mesmas da etapa de ajuste de parâmetros, essas URLs foram recolhidas no dia 24 de Novembro de 2014, enquanto as do ajuste foram obtidas em 2013.

A Tabela 7 apresenta os resultados obtidos individualmente por cada conjunto de características e pela combinação desses conjuntos, através de três tipos de

avaliação: Taxa de Precisão (TP), Taxa de Detecção (TD) e Falso Alarme (FA).

Tabela 7. Comparação entre Classificadores dos Grupos de Características

Classificadores	**J.48**			**KNN**		
Conjuntos	**TP**	**TD**	**FA**	**TP**	**TD**	**FA**
A	85,50%	85,22%	14,80%	**86,40%**	**85,98%**	14,00%
B	90,00%	89,95%	10,10%	**90,20%**	**90,11%**	9,90%
C	86,60%	86,52%	13,50%	**86,60%**	**86,55%**	13,40%
D	80,20%	79,20%	20,80%	**80,40%**	**79,49%**	20,50%
E	**87,70%**	**87,65%**	12,50%	87,30%	87,20%	12,80%
A+B+C	94,70%	94,67%	5,30%	**95,00%**	**94,97%**	5,00%
A+B+D	94,30%	94,32%	5,70%	**94,80%**	**94,77%**	5,20%
Todos	**95,00%**	**94,99%**	5,00%	94,90%	94,89%	5,10%
Classificadores	**SVM**			**Naive Bayes**		
Conjuntos	**TP**	**TD**	**FA**	**TP**	**TD**	**FA**
A	74,10%	74,10%	25,90%	74,30%	74,23%	25,80%
B	83,20%	82,64%	17,40%	79,50%	71,70%	28,30%
C	77,00%	75,90%	24,10%	73,30%	66,25%	33,70%
D	69,50%	69,22%	30,80%	70,20%	60,94%	39,10%
E	78,40%	77,31%	22,70%	72,80%	64,30%	35,60%
A+B+C	90,00%	89,95%	10,00%	80,00%	73,27%	26,70%
A+B+D	89,80%	89,83%	10,20%	75,30%	64,32%	35,70%
Todos	90,30%	90,30%	9,70%	75,30%	65,86%	34,10%

Comparando-se os resultados apresentados com os da Tabela 4, usada na comparação entre os classificadores para o ajuste dos parâmetros, percebe-se ainda que o melhor classificador geral (ou seja, para o grupo com todas as características) é o J.48. Contudo, na avaliação individual dos grupos por características, o classificador KNN foi o que apresentou os melhores resultados. Nas métricas de Taxa de Precisão (TP) e Taxa de Detecção (DT), ele obteve 86,40% e 85,98%, respectivamente, para o conjunto A; 90,20% e 90,11% para o conjunto B; 86,60% e 86,55% para o conjunto C e 80,40% e 79,49% para o conjunto D. Além disso, também obtive os melhores valores de TP e TD (95,00% e 94,97%; e 94,80% e 94,77%, respectivamente) para os conjuntos agrupados A+B+C e A+B+D. Já o classificador J.48 obteve os melhores resultados para o conjunto individual E (87,70% e 87,65%) e para o conjunto formado por todas as características (95,00% e 94,99%).

Um aspecto do classificador KNN que pode explicar essa melhor aderência aos dados testados diz respeito a normalização dos atributos. A normalização é usada para evitar que as medidas de distância utilizadas no cálculo do vizinho mais próximo sejam dominadas por um único atributo. E é justamente isso que acontece nesta base de dados, onde todos os atributos são inteiros positivos. Além disso, as características dos conjuntos A e B, intrinsicamente dependentes de serviços Internet (DNS, whois, popularidade), muitas vezes não retornam valores e, assim, são preenchidos com valores zero (0). Só para esclarecer melhor essa influência, as três características do conjunto A, *ip_associado*, *sn_associado* e *data_tempo_ativo*, obtiveram, respectivamente, 612, 4.584 e 7.328 valores sem resposta, substituídos por zero, ou seja, dos 60.000 valores esperados (20.000 de cada característica), 12.524 tiveram o zero atribuído. Para o conjunto B, as seis (6) características tiveram praticamente 1/3 de seus valores totais zerados (39.894 valores de um total de 120.000 esperados foram zerados).

De volta ao foco desta seção que é avaliar as hipóteses H3 e H4, fica claro que comparando as métricas de TP e TD do grupo A+B+C (95,00% e 94,97%) com as métricas de TP e TD do conjunto Total (95,00% e 94,99%), o uso de todas as

características não é relevante na obtenção dos melhores resultados. Mesmo que se argumente que os resultados do conjunto A+B+C são do classificador KNN e os do conjunto Total são do J.48, e por isso não devem ser comparados, ao se analisar os mesmos conjuntos (A+B+C e Total) somente entre os mesmos classificadores percebe-se que: no KNN, o conjunto A+B+C é melhor que o Total, e que no J.48, a diferença entre o Total e o A+B+C é de apenas 0,30%.

Assim, pode-se afirmar que a hipótese H3 é falsa e que a hipótese H4 é verdadeira. Em outras palavras, *não é necessário utilizar todas as características no processo de detecção de URLs maliciosas* e *os grupos de características permitem resultados adequados, e até melhores, no processo de detecção de URLs se comparados com características individuais.*

Diferenças nas bases de dados

Com a função de validar a hipótese H2, dois novos experimentos foram realizados. O primeiro deles avalia a relevância da característica para a análise, e o segundo, avalia o resultado dos quatro (4) classificadores em uma base diferente composta por URLs ligadas a blacklist.

Para avaliar a relevância de cada característica individualmente e, assim medir sua qualidade no processo de análise de uma URL, foi empregada a técnica de seleção de atributos, chamada Information Gain, disponível na ferramenta Weka como InfoGain. De modo geral, o Information Gain é um algoritmo de "ranking" fundamentado no conceito de entropia. No Weka, o InfoGain é independente de classificador. Valores altos no InfoGain significam que a característica é mais adequada no processo de predição. A Tabela 8 apresenta os resultados do InfoGain aplicado às 24 primeiras características mais relevantes nas bases DMOZ/PhishTank e DMOZ/-Blacklist.

Tabela 8. Comparação do InfoGain nas Bases DMOZ/PhishTank e DMOZ/Blacklist

DMOZ/PhishTank		DMOZ/Blacklist	
Característica	**InfoGain**	**Característica**	**InfoGain**
rank_google	0.390339	*qt_tok_url_barra*	0.70944
data_tempo_ativo	0.348147	*data_tempo_ativo*	0.452587
geo_localizacao	0.228261	*qt_tok_dir_barra*	0.276703
qt_tok_dom_ponto	0.158864	*geo_localizacao*	0.2612
qt_tok_url_barra	0.14602	*qt_tok_dom_ponto*	0.237529
qt_tok_dir_barra	0.118791	*comp_arquivo*	0.233644
comp_diretorio	0.115648	*rank_alexa*	0.180241
comp_url	0.105754	*comp_dominio*	0.176853
qt_tok_url_ponto	0.096124	*ip_assoc*	0.162585
rbl_check	0.082557	*qt_tok_url_ponto*	0.133253
comp_arquivo	0.074864	*rank_google*	0.120826
comp_dominio	0.073538	*presenca_marca*	0.102384
sn_assoc	0.068434	*sn_assoc*	0.067489
presenca_marca	0.056488	*rbl_check*	0.067109
rank_alexa	0.051651	*comp_diretorio*	0.06236
ip_assoc	0.038537	*comp_url*	0.052208
qt_tok_dir_ponto	0.035866	*comp_params*	0.037848
comp_params	0.029835	*qt_tok_arq_ponto*	0.032119
qt_tok_url_hifen	0.029591	*qt_tok_dir_hifen*	0.03101
qt_tok_dir_hifen	0.028234	*qt_tok_url_igualdade*	0.02972
qt_tok_url_ecomerc	0.024787	*qt_tok_par_igualdade*	0.029292
qt_tok_dom_hifen	0.02455	*qt_tok_arq_til*	0.022689
qt_tok_par_ecomerc	0.022983	*qt_tok_url_hifen*	0.018402
qt_params	0.022983	*qt_tok_url_interrog*	0.014041

Comparando-se os resultados do InfoGain para ambas as bases, nas 24

primeiras características, percebe-se claramente que existem grandes diferenças. A primeira delas diz respeito a quais são essas características. Na base DMOZ/PhishTank existem cinco (5) características (*qt_tok_dir_ponto*, *qt_tok_url_ecomerc*, *qt_tok_dom_hifen*, *qt_tok_par_ecomerc* e *qt_params*) que não estão presentes no InfoGain da outra base. Neste mesmo ponto de vista, na base DMOZ/Blacklist também existem cinco (5) características (*qt_tok_arq_ponto*, *qt_tok_url_igualdade*, *qt_tok_par_igualdade* , *qt_tok_arq_til* e *qt_tok_url_interrog*) que não aparecem no InfoGain da primeira base. Em comum, ambas as bases tem 19 características.

O segundo aspecto que as diferencia são os valores (importância) das características. Enquanto na base DMOZ/PhishTank, a característica mais relevante é *rank_google* com um valor de 0.390339 (39% de relevância), na base DMOZ/Blacklist, *qt_tok_url_barra* é a característica mais relevante com 0.70944 (70%). Já o segundo elemento em ambas as bases é *data_tempo_ativo*, mas sua relevância é maior na base DMOZ/Blacklist (0.452587) do que na base DMOZ/PhishTank (0.348147). Esse comportamento (maior relevância das características) é visto nos outros elementos da base DMOZ/Blacklist.

Desta forma, esta dissertação segue a posição que *a importância das características depende da base onde as URLs são coletadas.*

6. Conclusões e Trabalhos Futuros

Este trabalho apresentou uma investigação sobre a capacidade de validação e classificação de URLs como benignas e suspeitas/maliciosas através de determinadas características extraídas das próprias URLs, empregando técnicas de aprendizagem de máquina. Primeiramente, um estudo sobre as características extraíves de URLs foi realizado e uma taxonomia foi proposta. Em função dessa taxonomia, as principais características foram agrupadas e apresentadas. Em seguida, trabalhos relacionados foram apresentados visando mostrar os classificadores mais comuns empregados na classificação de URLs, bem como as características mais utilizadas. A extração das características ocorreu em grupos e foi descrita em termos de sua implementação.

Também foi apresentado o protocolo experimental envolvendo ajustes nos quatro (4) classificadores utilizados, bases de dados e outros dados. Para alcançar o objetivo proposto, vários experimentos foram realizados envolvendo URLs maliciosas e legítimas. Neste processo, foi imprescindível a elaboração de quatro (4) hipóteses para investigar a importância das características, de forma individual ou em grupo, e a relevância das características e da local (base) de onde as URLs são coletadas. Todas as hipóteses foram submetidas aos classificadores ou filtros e os resultados provam que:

1. O formato da URL possui certa influência na extração de características e, consequentemente, no processo de avaliação;
2. Não é necessário utilizar todas as características no processo de detecção de URLs maliciosas;
3. Os grupos de características permitem resultados adequados, e até melhores, no processo de detecção de URLs se comparados com características individuais;
4. A importância das características depende da base onde URLs são coletadas.

No que tange a trabalhos futuros, a descoberta e teste de novas características relevantes, bem como o uso de outros classificadores para obtenção de novos resultados são trabalhos a serem realizados.

Referências

Aaron and Rasmussen (2014). Global phishing survey: Trends and domain name use in 2h2013. `http://goo.gl/Fjkp9x`.

Akiyama, M., Yagi, T., and Itoh, M. (2011). Searching structural neighborhood of malicious urls to improve blacklisting. In *Proceedings of the 2011 IEEE/IPSJ International Symposium on Applications and the Internet*, pages 1–10. IEEE.

Anderson, D. S., Fleizach, C., Savage, S., and Voelker, G. M. (2007). Spamscatter: Characterizing internet scam hosting infrastructure. In *Proceedings of 16th USENIX Security Symposium on USENIX Security Symposium*, SS'07, pages 10:1–10:14, Berkeley, CA, USA. USENIX Association.

bin Lin, J. (2008). Anomaly Based Malicious URL Detection in Instant Messaging. Master's thesis, Dep. of Computer Science and Engineering, National Sun Yat-Sen University.

Canali, D., Cova, M., Vigna, G., and Kruegel, C. (2011). Prophiler: A fast filter for the large-scale detection of malicious web pages. In *Proceedings of the 20th International Conference on World Wide Web*, pages 197–206. ACM.

Choi, H., Zhu, B. B., and Lee, H. (2011). Detecting malicious web links and identifying their attack types. In *Proceedings of the 2Nd USENIX Conference on Web Application Development*, pages 11–11. USENIX.

Eshete, B., Villafiorita, A., and Weldemariam, K. (2013). Binspect: Holistic analysis and detection of malicious web pages. In *Security and Privacy in Communication Networks*, volume 106, pages 149–166. Springer Berlin Heidelberg.

Lee, B. T., Masinter, L., and Mccahill, M. (1994). RFC 1738: Uniform resource locator (URL). http://www.ietf.org/rfc/rfc1738.txt.

Ma, J., Saul, L. K., Savage, S., and Voelker, G. M. (2009). Beyond blacklists: Learning to detect malicious web sites from suspicious urls. In *Proceedings of the 15th ACM SIGKDD*, pages 1245–1254. ACM.

Maslennikov and Namestnikov (2012). Kaspersky security bulletin statistics 2012. `http://goo.gl/LfPhVD`.

Prakash, P., Kumar, M., Kompella, R. R., and Gupta, M. (2010). Phishnet: Predictive blacklisting to detect phishing attacks. In *Proceedings of the 29th Conference on Information Communications*, pages 346–350. IEEE.

Raymond, E. (1998). Book review: The essential perl books. *Linux J.*, 1998(46es).

Tsoumakas, G. and Vlahavas, I. (2007). Random k-labelsets: An ensemble method for multilabel classification. In *Proceedings of the 18th European Conference on Machine Learning*, pages 406–417. Springer-Verlag.

Zhang, J., Porras, P., and Ullrich, J. (2008). Highly predictive blacklisting. In *Proceedings of the 17th Conference on Security Symposium*, pages 107–122. USENIX.

Zhang, M.-L. and Zhou, Z.-H. (2007). Ml-knn: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038 – 2048.

