Xphide: Um Sistema Especialista para a Detecção de Phishing

Mateus L.S.D. de Barros[1], **Carlo M.R. da Silva**[2], **Péricles B.C. de Miranda**[1]

[1]DC – Universidade Federal Rural de Pernambuco (UFRPE)

[2]Campus Garanhuns – Universidade de Pernambuco (UPE)

mateuslins02@gmail.com, marcelo.revoredo@upe.br

pericles.miranda@ufrpe.br

Abstract. *Phishing is a type of cybercrime that targets the theft of a user's personal data through disguise and deception. This article proposes Xphide, a specialist system for detecting malicious pages. The basis of the system's construction was made through an in-depth analysis regarding relevant attributes for the description of web pages. This analysis served as input for the elaboration of Xphide's decision-making rules. The proposed system was evaluated in three different databases. The results showed that it surpassed traditional classification algorithms in terms of precision and recall, proving to be a promising alternative for the classification of web pages.*

Resumo. *Phishing é um tipo de crime cibernético que visa o roubo de dados pessoais do usuário por meios de disfarces e enganação. Este artigo propõe o Xphide, um sistema especialista para a detecção de páginas maliciosas. A base da construção do sistema foi feita através de uma análise aprofundada a respeito de atributos relevantes para descrição de páginas web. Esta análise serviu de insumo para a elaboração das regras do processo decisório do Xphide. O sistema proposto foi avaliado em três diferentes bases de dados. Os resultados mostraram que o mesmo superou algoritmos de classificação tradicionais em termos de precisão e revocação, se mostrando uma alternativa promissora para a classificação de página web.*

1. Introdução

Phishing se trata de ataques de engenharia social, que visam forjar entidades confiáveis, a fim de obter dados pessoais de suas vítimas [McGrath and Gupta 2008]. Com uma expansão que caminha junto com o avanço da internet, os *phishers*, como são conhecidos esse tipo de agente mal-intencionado, produzem cada vez mais novas formas enganação. Os sites phishing são domínios próprios destes criminosos, que utilizam diversas técnicas para simular um domínio legítimo, geralmente famoso. Dentre suas possíveis plataformas de distribuição, as comunicações eletrônicas, que englobam e-mails e mensagens de texto, mostraram-se as preferidas, graças ao grande leque de técnicas de enganação aptos a serem postos em prática nestes meios [Jagatic et al. 2007]. As mensagens enviadas pelos *phishers* tendem a ter links maliciosos para a inicialização de malware ou sites que forjam domínios famosos legítimos.

O phishing é considerado um crime cibernético [Mohammad et al. 2015]. Países como EUA, Inglaterra, Canadá e Austrália já aprovaram leis que criminalizavam esta

prática. Também foram realizadas campanhas de conscientização da população a respeito de medidas de prevenção aos ataques. Porém, o curto tempo de vida dos sites maliciosos impossibilita a apreensão dos *phishers*, visto que dificilmente são encontrados em fontes ainda ativas. Uma aliada a essa dificuldade é a chamada Curva de Aprendizado [Tsymbal 2004], conceito que consiste na capacidade do phishing de ser volátil. Novos meios de enganação estão sempre sendo desenvolvidos, fazendo com que as técnicas de detecção tornem-se obsoletas a longo prazo.

A fim de identificar páginas maliciosas, é preciso reconhecer padrões entre elas. Apesar da constante evolução do phishing, há certas práticas que estão frequentemente presentes, e ajudam na diferenciação de domínios seguros. Uma boa seleção de quais atributos utilizar para a classificação pode impactar positivamente na identificação dos sites maliciosos [Barros et al. 2019]. Desta forma, é possível a extração de informações da página alvo, orientadas pelos atributos, para chegar a um consenso sobre sua legitimidade. Tendo visto esta dinâmica, torna-se necessário um sistema que possa classificar a página, a partir de atributos extraídos da mesma.

O presente trabalho propõe um sistema especialista para a detecção de páginas maliciosas, chamado de Xphide. O Xphide é baseado em regras, criadas a partir de um estudo aprofundado acerca de atributos propostos em [Barros et al. 2019]. Deste modo, foi feita uma análise comparativa dos atributos escolhidos, o que permitiu a distribuição de pesos por ordem de relevância. O Xphide foi comparado com algoritmos clássicos de aprendizagem de máquina em termos de acurácia, precisão e *recall*. A abordagens foram avaliadas em bases do Phishtank[1] e Openphish[2]. Os resultados mostraram que o Xphide obteve resultados superiores aos dos algoritmos clássicos, mostrando-se ser uma alternativa promissora para a classificação automática de páginas maliciosas.

Este artigo está dividido na seguinte forma: a Seção 2 apresenta o referencial teórico sobre a detecção de phishing; a Seção 3 apresenta os trabalhos relacionados à detecção de phishing através de inteligência artificial. A Seção 4 apresenta a proposta adotada para o desenvolvimento do sistema especialista, assim como a explicação detalhada de cada atributo extraído; na Seção 5 detalha a metodologia adotada para a execução do experimento. Na Seção 6, são apresentados os resultados obtidos; e na Seção 7, a conclusão e trabalhos futuros.

2. Problema de Detecção de Phishing

O phishing é um problema antigo da computação, cujo nascimento e desenvolvimento caminha junto com a popularização da internet. Em [Banu and Banu 2013], são exemplificados alguns tipos de phishing que surgiram juntos à evolução da tecnologia. O mais comum é conhecido como Clone Phishing, que se passam por sites de marcas famosas para obter dados pessoais do usuário. Spear Phishing apresenta similaridade com o anterior, porém visando um grupo específico de vítimas. DNS-Based Phishing, também conhecido como *Pharming*, mapeia um nome de domínio de um site legítimo em um endereço IP de um site malicioso. O ataque Man-In-The-Middle opera diferencialmente aos tipos anteriores, ocorrendo quando o *phisher* intercepta uma mensagem de uma empresa oficial para um usuário, modificando-a e colocando seu link de domínio malicioso.

[1]https://www.phishtank.com/

[2]https://openphish.com/

Em [Fette et al. 2007], é apresentado alguns dos primeiros métodos para a contenção deste problema. O primeiro foi através de *toolbars* em navegadores. Apesar de apresentar uma acurácia de 85%, este método apresentava duas desvantagens fundamentais. A primeira seria uma redução na quantidade de informação textual fornecida pelos e-mails analisados. A segunda seria a incapacidade de fornecer uma proteção segura, pois o acesso não era de fato bloqueado, podendo ser ignorado pelo usuário.

O segundo método caracterizou-se pela filtragem de e-mails. Embora sendo mais robusto em comparação ao anterior, ainda apresentava falhas nítidas. Uma vez que limitava-se à linguagem e filtro do texto das mensagens, com uma análise superficial da URL. Esta abordagem gerava uma alta quantidade de falsos negativos, e mostrava-se suscetível à problemática da Curva de Aprendizado [Tsymbal 2004] voltada ao phishing.

Diferentes trabalhos perceberam a presença de padrões entre os phishings, sendo possível, portanto, o uso de algoritmos inteligentes para a automatização da classificação de páginas web. A seguir, são detalhados alguns dos principais trabalhos que utilizaram inteligência artificial para a classificação de páginas maliciosas.

3. Trabalhos Relacionados

A detecção de phishing é um problema que vem sendo cada vez mais relevante e estudado. Ao longo do tempo, diferentes atributos foram identificados e usados para caracterizar phishing [Banu and Banu 2013]. Com isso, diferentes bases de dados foram construídas para auxiliar especialistas no treinamento e teste de modelos de classificação. Em [Abdelhamid et al. 2014], são apresentados alguns dos atributos mais populares em bases de dados de acesso público, e em pesquisas relacionadas. Atributos como *long URL*, que verifica o tamanho completo do link; *having @ symbol*, que verifica a presença do caractere '@' na URL; ou *adding Prefix and suffix*, que analisa a presença de prefixos ou sufixos no domínio; fazem parte da categoria voltada à análise da URL. Também estão presentes atributos voltados ao corpo da página, como *Pop-up window*, e atributos com uso de terceiros, como *age of domain*. Alguns destes estão presentes neste trabalho, e serão detalhados ao longo do artigo.

Em [Moghimi and Varjani 2016], é utilizado um algoritmo de classificação de páginas maliciosas utilizando diferentes atributos, como: *IP address*, que verificava a presença de endereços IP no lugar do domínio; ou *SSL certificate*, observando a presença do certificado SSL na página analisada; e atributos que utilizam *greylist*. Este método consiste na criação de uma lista de palavras pré-estabelecidas para serem comparadas com elementos da URL analisada. A utilização de *greylists* e *blacklists*, também é utilizada em alguns atributos deste trabalho, e será melhor aprofundada posteriormente neste artigo.

Em [Barros et al. 2019], é realizada uma seleção de atributos utilizados na detecção de phishing, bem como uma comparação entre diferentes algoritmos de classificação presentes na literatura. Os algoritmos escolhidos foram *Support Vector Machines* (SVM), *Decision Trees* (DT) e *Neural Networks* (NN). Os resultados revelaram uma superioridade de DT nas taxas de acurácia e *F1-Score*. A seleção de atributos foi capaz de diminuir uma bases de 30 para somente 12 atributos, mantendo um alto nível das mesmas taxas mencionadas nos resultados de DT. Também foram destacados *SSLfinal_State* e *URL_of_Anchor* como os atributos mais relevantes na classificação.

Como mencionado em [Banu and Banu 2013], diferentes tipos de phishing reque-

rem diferentes abordagens. O Spear Phishing, mais especificamente, possui uma maior dificuldade em ser identificado, por se tratar de um ataque direcionado. Portanto, algumas estratégias podem ser destacadas para explicar sua atuação [Silva et al. 2019]. São elas:

- Fidedignidade, que descreve a alta ou baixa riqueza de detalhes da fraude em relação à página genuína, sendo uma estratégia bem variável.
- Ofuscação, que descreve as investidas do fraudador em ocultar informações que poderiam ser visíveis ao usuário final, mas devido a quantidade alta ou baixa de caracteres, alguns detalhes podem não ser observados.
- Propagação, que descreve alguns comportamentos que visam aumentar o alcance das fraudes em um grande número de usuários, como burlar mecanismos baseando em listas negras.
- Sazonalidade, que descreve a sensibilidade do phishing à eventos do calendário.
- Volatilidade, que remete ao tempo de vida curto do mesmo, evidenciando que a fraude é rapidamente abandonada por seu criador.

Algumas destas estratégias foram levadas em consideração na criação do sistema deste trabalho, e será discutida na seção seguinte. Os trabalhos supracitados utilizam classificadores clássicos e atributos sugeridos por bases públicas para classificação de sites maliciosos. Há, porém, uma carência de trabalhos que utilizem o potencial de sistemas especialistas na resolução do problema, que avaliam os atributos existentes, ponderando-os por relevância. Diante disso, este trabalho propõe o Xphide, um sistema especialista capaz de extrair atributos pré-definidos em tempo real, e classificar a página analisada entre segura ou maliciosa, aliada a boas taxas de predição. Na seção a seguir será detalhada a construção do Xphide, desde a análise de atributos relevantes e seus pesos na classificação, até a elaboração das regras do processo decisório.

4. Xphide: Um Sistema Especialista para a Detecção de Phishing

O Xphide é um sistema especialista baseado em regras. As regras que guiam o processo decisório foram definidas através de um estudo aprofundado sobre os atributos provenientes de três bases de dados: duas bases com exemplos de phishing válidos (Phishtank e Openphish), e outra base de exemplos de não phishing (Phishtank), também chamados de phishing inválidos. Estas bases foram selecionadas, pois são robustas e largamente utilizadas na literatura [Banu and Banu 2013, Barros et al. 2019].

A seção 4.1 apresenta o estudo realizado sobre os atributos e exemplos contidos nas bases de dados selecionadas, para a construção de regras; a seção 4.2 apresenta as regras definidas; e a seção 4.3 apresenta o funcionamento do Xphide.

4.1. Análise de Características

Para a elaboração das regras do Xphide, foi realizada uma análise qualitativa a respeito dos atributos mais relevantes para a diferenciação entre as classes phishing e não phishing. Foram analisadas três categorias de atributos: atributos estáticos instantâneos, extraídos automaticamente da URL; atributos estáticos de lista, extraídos da URL com auxílio de listas criadas por nós, oriundas de diferentes fontes para suprir as individualidades de cada um; e atributos dinâmicos, extraídos com auxílio de terceiros, sendo estes componentes não exclusivos da URL. Diferentemente dos estáticos, eles dependem da conexão do usuário para serem extraídos. Atributos que contêm comparação de quantidade ou

fração de tempo foram sintetizados a partir de médias resultantes nas bases de dados selecionadas.

Com o intuito de produzir classificações que privilegiem atributos de maior relevância, foi elaborada uma estratégia de pontuação por atributo. Para cada atributo atribuímos um valor base +1. Levando em consideração as conclusões obtidas em [Silva et al. 2019], definimos que os atributos relacionados à estratégia de Ofuscação, em sua maioria, obtém maior peso (+3), por conta da maior capacidade de enganação por meio dos *phishers*. Em outros casos, alguns atributos voltados às estratégias de propagação também recebem um maior destaque (+3). Os relacionados à estratégia de volatilidade, por sua vez, permanecem com peso menor (+1). O motivo se dá pelo fato da análise por base de tempo ser instável e totalmente dependente de bases frequentemente atualizadas. Como forma de aprimorar esta classificação, também contabilizamos a taxa de ocorrência dos atributos nas bases selecionadas. Isto permitiu o estudo de casos onde possam surgir exceções à regra anteriormente estabelecida.

Os atributos pertencentes à categoria estática instantânea podem ser vistos na Tabela 1, assim como suas descrições, pontuações e relevâncias. Suas taxas de ocorrência podem ser vistas na Figura 1.

Atributos	**Descrição**	**Pontuação**	**Relevância**
Tem Longa URL	Caso a URL possua mais de 60 caracteres, é considerada longa	+1	Baixa
Tem Longo Domínio	Caso o domínio possua entre 11 e 20 caracteres, é considerado longo. Se possuir mais que 20 caracteres, é considerado muito longo	+1 ou +3	Média
Tem Longo Subdomínio	Caso o subdomínio possua entre 9 e 16 caracteres, é considerado longo. Se possuir mais que 16 caracteres, é considerado muito longo	+1 ou +3	Média
Quantidade suspeita de subdomínios	Caso a quantidade de subdomínios seja entre 2 e 3, é considerado suspeito. Se for maior que 3, é considerado malicioso	+1 ou +3	Média
Quantidade suspeita de *path*	Caso a quantidade de *paths* seja maior que 4, é considerado suspeito	+1	Baixa
Aparenta redirecionamento	Caso seja identificada a presença de uma URL dentro do *path* ou *query*, é considerado suspeito	+1	Baixa
Tem prefixo ou sufixo	Caso seja identificada a presença do caractere hífen (-) no domínio ou subdomínio, é considerado suspeito	+1	Baixa
Tem @	Caso seja identificada a presença de @ na URL, é considerado suspeito	+1	Baixa
Tem URL encoded	Caso seja identificada a presença de %, seguido de dois dígitos hexadecimais, é considerado malicioso	+3	Alta

Tabela 1. Atributos estáticos instantâneos

Dos atributos que compõem esta categoria, "Tem URL encoded"destaca-se por ter a maior capacidade de ofuscação. Isto se dá por ele permitir a ocultação de elementos da URL, mascarando-os com dígitos hexadecimais. É válido destacar também sua presença totalmente nula entre os phishing inválidos, tornando-o um atributo importante para a detecção dos phishings válidos. Por esta razão, recebeu uma pontuação +3.

Os cinco primeiros atributos da Tabela 1, relacionados a tamanho ou quantidade, possuem a capacidade de serem determinantes para a classificação em casos exorbitantes. Porém, somente "Tem longo domínio", "Tem longo subdomínio"e "Quantidade suspeita

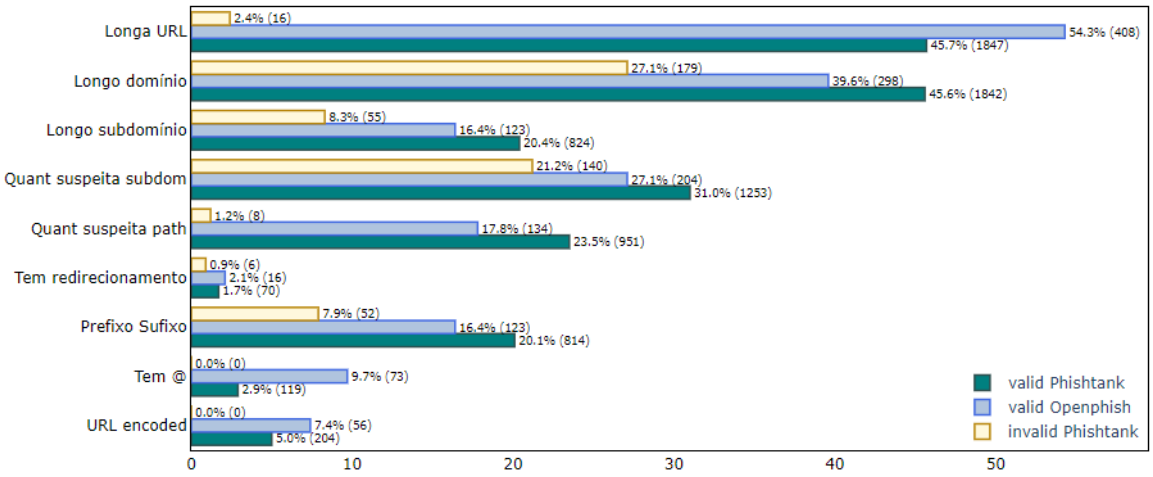

Figura 1. Ocorrência dos atributos estáticos instantâneos em cada uma das bases.

de subdomínios"têm a chance de alcançarem a pontuação +3 nestes cenários. Apesar da grande diferença de ocorrências nos atributos "Tem Longa URL"e "Quantidade suspeita de path", eles não podem ser considerados atributos de alta relevância. Isto se dá pelo fato que em sistemas reais (que não são phishing) relacionados a login ou comércio, as *query strings* podem tornar-se consideravelmente extensas, influenciando, portanto, na taxa de falsos positivos. Por este motivo, receberam um peso menor (+1).

A Tabela 2 apresenta os atributos pertencentes à categoria estática de lista, bem como suas definições, pontuações e relevâncias. A Figura 2 mostra suas taxas de ocorrência nas bases.

Atributos	**Descrição**	**Pontuação**	**Relevância**
Quantidade suspeita de separadores	Caso a quantidade de separadores contabilizados seja superior a 10, é considerado suspeito. A lista foi obtida a partir das colunas *Unreserved* e *Reserved* do domínio[3].	+1	Baixa
Tem Referência a formulário	Criamos uma lista composta por 61 palavras relacionadas a *login*, bancos e *e-commerce*. Caso seja identificada a presença de algum elemento da lista, é considerado suspeito	+1	Baixa
Tem TLD muito explorado	Contabilizamos os 10 *top-level domains* mais utilizados entre os phishing válidos. Caso seja identificada a presença de um deles, é considerado suspeito	+1	Baixa
Tem palavra homográfica	Criamos uma lista composta por nomes de marcas famosas. Caso seja identificado algum elemento da lista no subdomínio ou no *path* da URL, é considerado malicioso	+3	Alta
Tem subdomínio forjando TLD	Caso seja identificada a presença de *top-level domains* no subdomínio, isolado ou como prefixo ou sufixo, é considerado malicioso	+3	Alta
Tem host na *greylist*	Contabilizamos os 10 domínios mais utilizados entre os phishing válidos. Caso seja identificada a presença de um deles, é considerado malicioso	+3	Alta

Tabela 2. Atributos estáticos de lista

Os 3 últimos atributos desta categoria, "Tem palavra homográfica", "Tem subdomínio forjando TLD"e "Tem host na greylist", são fortemente ligados à estratégia de

[3]Bases de dados utilizadas: https://developers.google.com/maps/documentation/urls/url-encoding

propagação. Seu uso é frequentemente atrelado à presença de marcas ou domínios seguros em meio à URL maliciosa, com o objetivo de levar as vítimas a pensar se tratar de um website confiável. Eles mantêm uma baixa presença entre os phishing inválidos, sendo importantes na distinção entre as classes positiva e negativa. Desta forma, recebem um alto nível de relevância (+3). Apesar do atributo "Tem TLD muito explorado"também utilizar desta estratégia, sua taxa de ocorrência é ainda maior entre os phishing inválidos, o que o torna inviável de receber um peso maior. Mesmo com a grande diferença de aparição entre as bases, "Quantidade suspeita de separadores"cai na mesma questão dos atributos "Tem longa URL"e "Quantidade suspeita de path", da categoria anterior. Para evitar falsos positivos, foi atribuído um peso menor a este atributo.

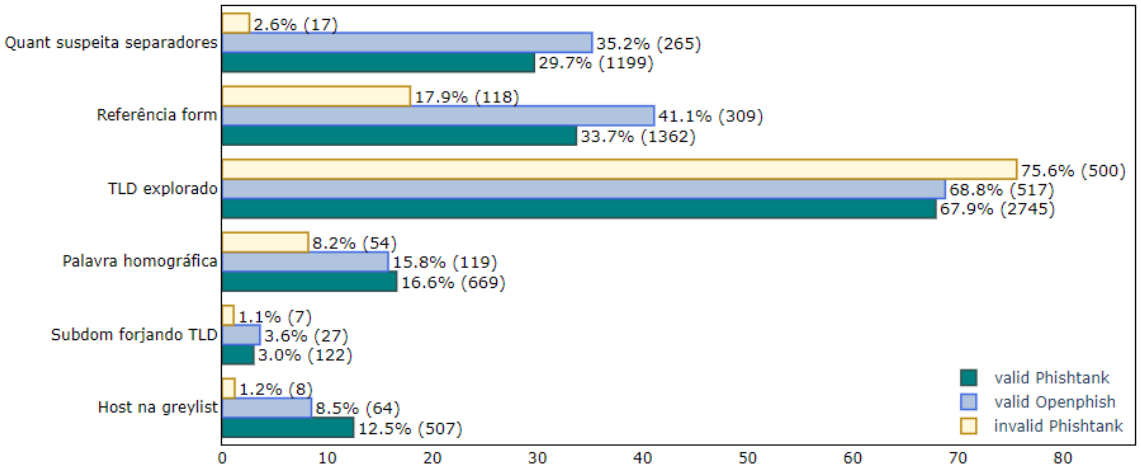

Figura 2. Ocorrência dos atributos estáticos de lista em cada uma das bases.

Por fim, a Tabela 3 mostra os 3 atributos pertencentes à categoria dinâmica, junto com suas descrições, pontuações e relevâncias. A Figura 3 mostra suas taxas de ocorrência. É importante salientar que os atributos que possuem intervalos de tempo estão sujeitos à base de dados nos quais são treinados. Diferentes bases podem fornecer diferentes datas que influenciam na média obtida.

Atributos	**Descrição**	**Pontuação**	**Relevância**
Tem certificado SSL inválido	É verificada a existência de um certificado SSL válido na página analisada. Caso não seja identificada ou esteja inválido, é considerado malicioso	+3	Alta
Tem SSL recentemente validado	Caso possua SSL válido, é verificado se ele foi validado recentemente. Caso tenha sido em um período anterior a 125 dias, é considerado suspeito	+1	Baixa
Tem domínio recentemente registrado	Através do *Whois*, extraímos última data de registro do domínio e calculamos a diferença de dias. Caso o resultado seja inferior a 2383 dias, é considerado suspeito	+1	Baixa

Tabela 3. Atributos Dinâmicos

O atributo "Tem SSL inválido"foi considerado de alta relevância, recebendo +3 de pontuação. Isso se explica pois o mesmo é presente na grande maioria dos sites minimamente confiáveis. Logo, por motivos de precaução, é mais seguro evitar o acesso a um domínio que não possua um SSL, justificando sua alta relevância. "Tem SSL recentemente validado"e "Domínio recentemente registrado"permanecem com pesos padrão por

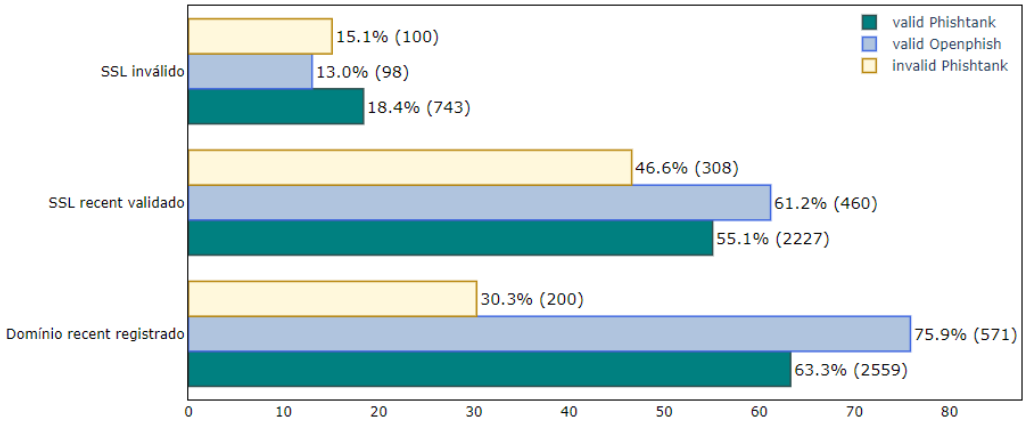

Figura 3. Ocorrência dos atributos estáticos dinâmicos em cada uma das bases.

utilizarem técnicas de volatilidade, sendo suscetíveis a constantes mudanças a variar das bases de dados adotadas.

4.2. Regras Elaboradas

Após a análise dos atributos, e a ponderação dos mesmos baseada na relevância, foi possível a elaboração de regras para a classificação. Foram construídos três modelos (diagramas de regras), cada um especializado em uma categoria de atributos. Logo, existe o modelo que leva em consideração atributos estáticos instantâneos (MEI), um modelo de atributos estáticos de lista (MEL), e um modelo de atributos dinâmicos (MD). As seções a seguir apresentam as regras definidas em cada um dos modelos.

4.2.1. Modelo de Atributos Estáticos Instantâneos

Este modelo leva em consideração apenas os atributos estáticos instantâneos. O diagrama de regras deste modelo pode ser visto na Figura 4.

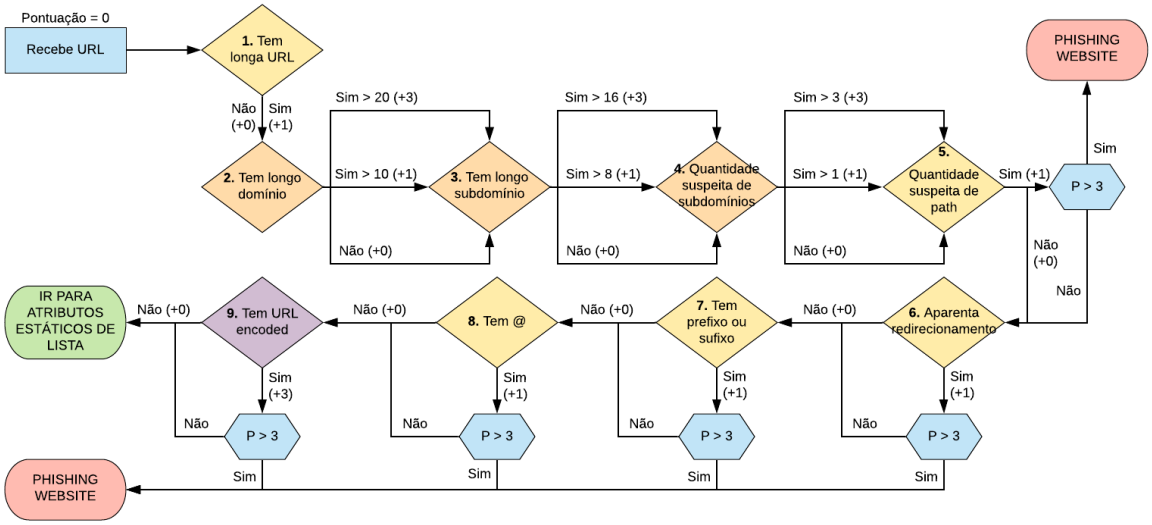

Figura 4. Modelo da sub-etapa estática instantânea.

Para uma dada URL, verifica-se, de forma encadeada, as características da mesma. Cada losango no diagrama se trata de uma estrutura condicional, e dependendo da resposta para a condição corrente atribui-se soma-se uma pontuação à nota (P) da URL. No MEI, existem alguns pontos de gatilho (hexágonos) que verificam se a pontuação ultrapassou o valor três. Caso $P > 3$ em alguns destes gatilhos, a URL é considerada phishing. Caso P nunca atinja o valor três, o processo continua para o próximo modelo, o MEL, que será detalhado a seguir.

4.2.2. Modelo de Atributos Estáticos de Lista

O MEL possui um conjunto de regras que leva em consideração apenas os atributos estáticos de lista (ver Figura 5). O MEL só é executado se nenhum gatilho for ativado no MEI. Além disso, a nota da URL (P), após o MEL ser iniciado, é zerada. O processo se dá de forma similar ao apresentado no MEI, existem estruturas condicionais encadeadas, e dependendo das respostas são somados pontos à P. Neste modelo existe um único gatilho, que verifica se $P > 2$. Em caso positivo, a URL é considerada phishing. Além desse gatilho, existem três verificações que podem definir a classe como sendo phishing: se tem palavra homográfica, se te subdomínio forjando TLD, ou se tem host na *greylist*. Se uma delas for positiva, é somado o valor três a P, e a URL é considerada phishing. Por outro lado, se o valor de P, durante todo o diagrama, nunca atingir valores superiores a dois, o processo continua para próximo e último modelo, o MD.

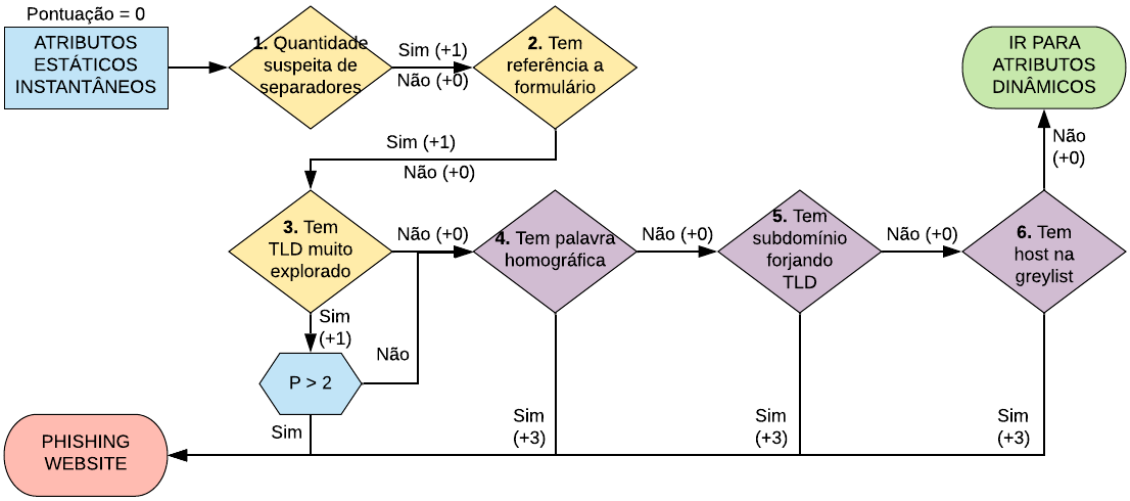

Figura 5. Modelo da sub-etapa estática de lista.

4.2.3. Modelo de Atributos Dinâmicos

O MD possui um conjunto de regras que leva em consideração apenas os atributos dinâmicos. O diagrama de funcionamento do MD pode ser visto na Figura 6. O MD só é executado se nenhuma condição ou gatilho levar à classificação de *phishing*. Assim como no modelo anterior, a nota da URL (P), é zerada. Logo no início do MD verifica-se a validade do certificado SSL. Se a resposta for negativa, soma-se três à P e a URL é definida como phishing. Caso contrário, as verificações continuam. Neste modelo existe um

único gatilho, que verifica se $P > 1$. Em caso positivo, a URL é considerada phishing. Se o valor de P nunca atingir valores superiores a um, a URL é considerada segura e o processo decisório é encerrado.

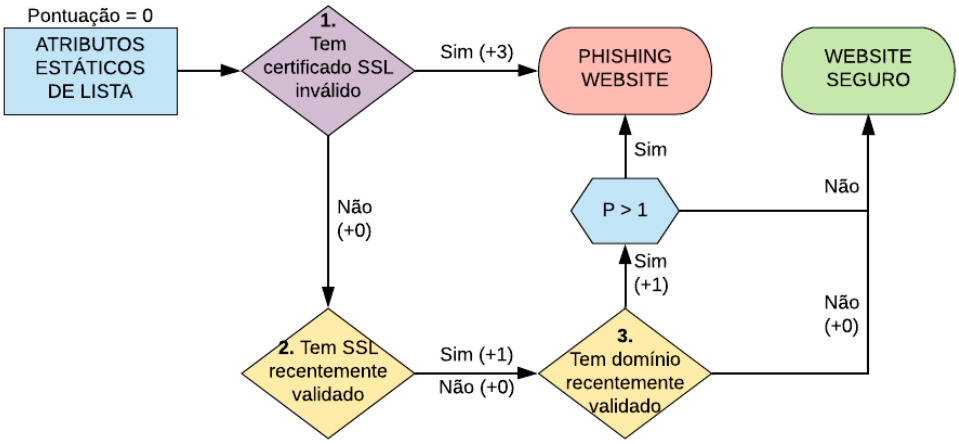

Figura 6. Modelo da sub-etapa dinâmica.

4.3. Funcionamento do Xphide

A Figura 7 apresenta o *pipeline* de funcionamento do Xphide. Para iniciar a classificação, o sistema recebe a URL a ser investigada, e realiza uma requisição para o servidor. Se for identificado que o website está *offline*, o processo é interrompido. Caso contrário, o processo segue para a etapa de extração de características. Nesta etapa, os valores dos atributos, discutidos na seção 4.1, são extraídos. Em seguida, as características são passadas para o *core* do Xphide, seu conjunto de regras. Os modelos MEI, MEL e MD são executados seguindo o procedimento detalhado na seção 4.2, produzindo uma classe resultante: phishing ou não phishing (website seguro).

O Xphide apresenta dois aspectos positivos. O primeiro é que suas regras foram produzidas manualmente através de um estudo profundo sobre a relevância dos atributos. O segundo aspecto é que o processo de classificação do Xphide pode não precisar avaliar todos os atributos da URL, uma vez que tem a capacidade de interromper seu processo caso provas suficientes tenham sido identificadas para a classificação. Este segundo aspecto pode tornar o Xphide mais eficiente que classificadores clássicos.

5. Metologia Experimental

Neste trabalho, foi realizado um experimento quantitativo, que analisa os resultados obtidos na classificação em cada uma das bases de dados selecionadas. A proposta é avaliada em termos de acurácia, precisão e *recall*, e é comparada com a *Support Vector Machines* (SVM) e *Random Forest* (RF), classificadores clássicos largamente utilizados.

5.1. Bases de Dados

Utilizamos 3 bases de dados de diferentes repositórios. Realizamos um tratamento para reduzir a quantidade de domínios repetidos, assim como reduzir possíveis casos de falsos positivos. Como resultado obtivemos duas bases de phishing válidos, uma do Phishtank contendo 4040 amostras, e uma do Openphish composta por 752 amostras. Também

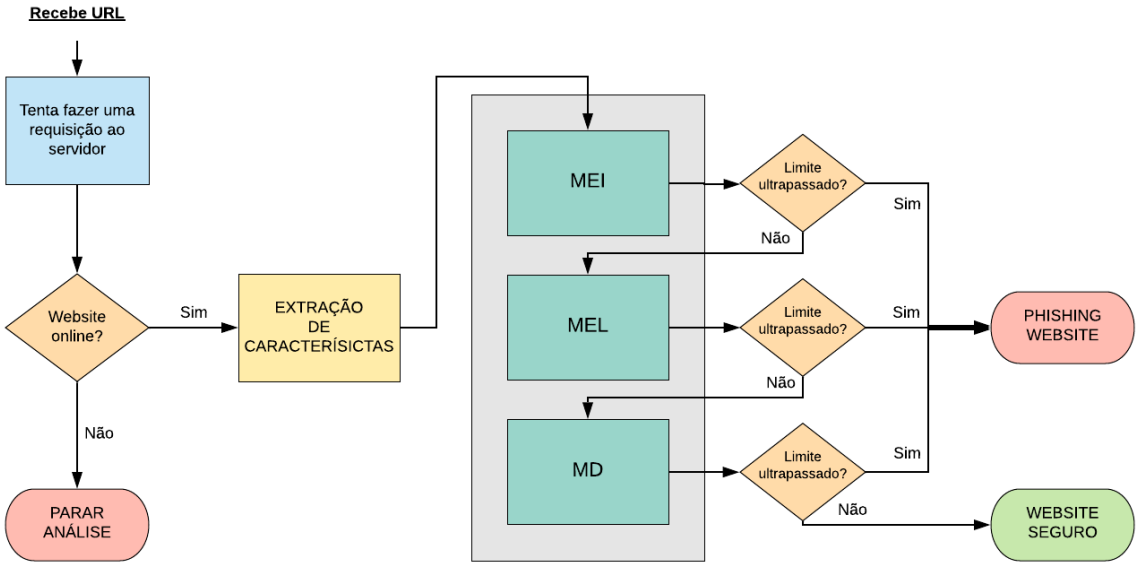

Figura 7. *Pipeline* de funcionamento do Xphide.

obtivemos uma base de phishing inválidos, websites que aparentam serem maliciosos mas provam-se seguros depois de testes, do Phishtank, contendo 661 amostras. Todas as bases, com seus devidos tratamentos, estão disponíveis publicamente[4].

5.2. Medidas de Avaliação

As medidas utilizadas para avaliar o Xphide se tratam da acurácia, precisão e *recall*.

A acurácia se trata da fração de classificações que o modelo acertou. Ela é calculada da seguinte forma:

$$Acurácia = \frac{TP + TN}{TP + TN + FP + FN}. \tag{1}$$

Vale salientar que uma acurácia elevada indica que o classificador acertou muitos exemplos na base de dados. No entanto, não se sabe ao certo o número de acertos por classe. Por este motivo, a análise da acurácia é comumente acompanhada pela análise da precisão e da revocação.

A precisão é a proporção de exemplos da classe positiva (é phishing) que foram classificadas corretamente. Ela é calculada da seguinte forma:

$$Precisão = \frac{TP}{TP + FP}. \tag{2}$$

O *recall* é a proporção de exemplos da classe negativa (não é phishing) que foram classificadas corretamente. Ela é calculada da seguinte forma:

$$Recall = \frac{TP}{TP + FN}. \tag{3}$$

[4]Bases de dados utilizadas: https://www.dropbox.com/s/b3qd1kus21fhh1c/samples.7z?dl=0

Nas equações, TP é o número de verdadeiros positivos, ou seja, phishings classificados corretamente; TN é o número de verdadeiros negativos, ou seja, não phishings classificados corretamente; FP é o número de falsos positivos, ou seja, URLs classificadas como phishings, não sendo; e FN é o número de falsos negativos, ou seja, URLs classificadas como não phishings, sendo phishing.

No contexto deste trabalho, a precisão e *recall* são medidas de avaliação muito importantes, pois não deseja-se permitir que o usuário acesse URLs maliciosas (FN), nem deixe de acessar URLs não maliciosas (FP). Altos valores de precisão e *recall* significam que o classificador sabe diferenciar bem a classe phishing da não phishing.

5.3. Metodologia de Avaliação

O sistema especialista proposto foi avaliado em termos de acurácia, precisão e *recall*, e comparado com dois algoritmos clássicos de aprendizagem de máquina, a *Support Vector Machines* (SVM) e *Random Forest* (RF). A SVM e o RF foram avaliados através do experimento de validação cruzada com 10 folds. Para as medidas de avaliação serem calculadas, as bases de phishings válidos e inválidos foram concatenadas, formando uma única base de dados. Já o Xphide possui suas regras pré-definidas, sem a necessidade de treinamento. Por este motivo, não foi avaliado através da validação cruzada assim como os algoritmos clássicos. Para ser avaliado, o Xphide precisou ser executado em cada base de dados, sendo sua acurácia, precisão e *recall* calculados à posteriori.

5.4. Recursos de Hardware e Software

Os algoritmos de classificação utilizados são provenientes da biblioteca *Scikit-Learn*[5]. Todos os testes foram feitos em um computador Intel Core i3-7100 CPU @ 3.90 GHz, com memória RAM de 8 GB.

6. Resultados

A base do Phishtank é composta por 4040 registros de sites considerados phishing (phishings válidos), não havendo instâncias do tipo não phishing. O Xphide foi avaliado nesta base, e como se pode ver na Figura 8-(Esq.), das 4040 URLs maliciosas, 3.697 foram classificadas corretamente como phishing. As demais 342 URLs, foram classificadas incorretamente como não phishing. O Xphide também foi avaliado na base de dados OpenPhish, composta por 752 URLs maliciosas. Como se pode ver na Figura 8-(Centro), o sistema especialista classificou corretamente 717 URLs como sendo maliciosas, errando em apenas 35 exemplos.

Adicionalmente, o Xphide também foi avaliado na base de phishings inválidos, URLs suspeitas mas que não são phishing (ver Figura 8-(Dir.)). Das 661 URLs totais contidas na base, 382 foram classificadas corretamente como não phishing. Os outros 279 exemplos foram classificados incorretamente como phishing.

Diante de todos os resultados apresentados, a seguir serão apresentados os resultados do Xphide em termos de acurácia, precisão e *recall*. O $TP = 4.414$, sendo 3.697 acertos de phishings válidos na Phishtank, somados aos 717 do Openphish. O $TN = 382$, os 382 exemplos corretamente classificados como não phishings, da base de

[5]https://scikit-learn.org

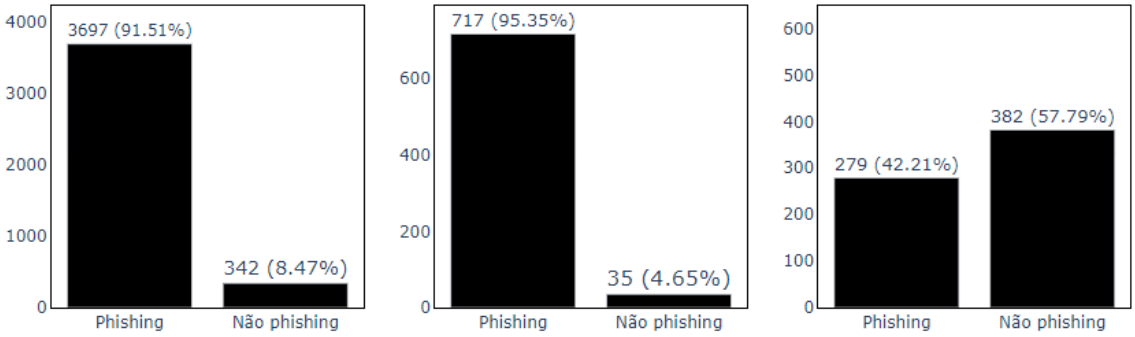

Figura 8. Esquerda: Resultados da classificação na base de phishings válidos do Phishtank. Centro: Resultados da classificação na base phishings válidos do Openphish. Direita: Resultados da classificação na base de phishings inválidos do Phishtank

dados de phishings inválidos do Openphish. $FP = 279$, os 279 exemplos incorretamente classificados como phishings, provenientes da base de phishings inválidos do Openphish. $FN = 377$, sendo 342 classificadas incorretamente como não phishing da base do Phishtank, somados aos 35 exemplos classificados incorretamente do Openphish. Com isso, temos uma acurácia igual a $87,96\%$; precisão igual a 94%; e *recall* igual a $92,13\%$.

A fim de complementar o experimento quantitativo, aplicamos a SVM e o RF na base de dados unificada (Phishtank válidos + Phishtank inválidos + Openphish válidos). Todos os algoritmos deste experimento foram utilizados com suas configurações padrão disponíveis na biblioteca *Scikit-Learn*. Os resultados obtidos pela SVM e RF são comparados com os do Xphide, e podem ser vistos na Tabela 4.

Algoritmo	**Acurácia**	**Precisão**	***Recall***
SVM	**91,96%**	69,51%	62,17
RF	91,72%	69,64%	56,41
Xphide	87,96%	**94%**	**92,13%**

Tabela 4. Acurácia, precisão e *recall* obtidas pelas diferentes abordagens.

Como se poder ver, o Xphide alcançou bom desempenho em todas as medidas de avaliação. Em termos de acurácia, seu desempenho também foi satisfatório. O Xphide alcançou $87,96\%$, enquanto a SVM e o RF alcançaram $91,96\%$ e $91,72\%$, respectivamente. Embora, o Xphide tenha alcançado um resultado de acurácia inferior aos da SVM e RF, o resultado da acurácia precisa ser avaliado em conjunto com a precisão e a *recall*, principalmente em um cenário de desbalanceamento de classes. Com relação à precisão e *recall*, o Xphide apresentou resultados muito superiores aos da SVM e do RF. Isso mostra que a nossa abordagem foi capaz de evitar um número elevado de FN e FP, conseguindo diferenciar as classes envolvidas. Os baixos valores de precisão e *recall* da SVM e RF, mostram que eles acertaram muitos exemplos (acurácia alta), mas a maioria dos exemplos classificados corretamente pertencem à classe positiva, que é a majoritária. Como a base de dados é desbalanceada, tanto a SVM quanto a RF erraram quase todos os exemplos da classe minoritária (classe negativa). Deste forma, o Xphide se mostrou uma solução promissora para a detecção de páginas maliciosas, mesmo em um conjunto de dados desbalanceado.

7. Conclusão

A tarefa de detectar Phishing não é trivial, pois envolve o entendimento sobre quais atributos são relevantes e descritivos para a classificação. Além disso, bases de dados com exemplos de páginas web, maliciosas ou não, são naturalmente desbalanceadas, dificultando a tarefa de classificadores clássicos.

Este trabalho propôs o Xphide, um sistema especialista para a detecção de páginas maliciosas. O Xphide teve o seu conjunto de regras elaborado a partir de um estudo aprofundado sobre a relevância de atributos presentes na literatura, direcionando assim o processo decisório. O Xphide foi avaliado em bases de dados do Phishtank e Openphish com exemplos positivos e negativos de phishing. Os resultados mostraram que o Xphide alcançou um valor de acurácia competitivo em relação à algoritmos de classificação clássicos. No entanto, em termos de precisão e *recall*, o Xphide alcançou resultados muito superiores, mostrando-se capaz de distinguir as classes, com baixas taxas de falsos positivos e negativos.

Como trabalhos futuros, pretende-se aplicar o Xphide a um portal online de execução em tempo real, a fim de classificar páginas maliciosas. Também é válido mencionar o uso de processamento de linguagem natural para extração de características do código da página (HTML, JavaScript). O intuito é verificar se há informação relevante para a distinção entre as classes, visando reduzir ainda mais a presença de FP e FN.

Referências

Abdelhamid, N., Ayesh, A., and Thabtah, F. (2014). Phishing detection based associative classification data mining. *Expert Systems with Applications*, 41(13):5948–5959.

Banu, M. N. and Banu, S. M. (2013). A comprehensive study of phishing attacks. *International Journal of Computer Science and Information Technologies*, 4(6):783–786.

Barros, M., Silva, C., and Miranda, P. (2019). Adoção da seleção de características como mecanismo antiphishing: aplicabilidade e impactos. In *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional*, pages 214–225. SBC.

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

Jagatic, T. N., Johnson, N. A., Jakobsson, M., and Menczer, F. (2007). Social phishing. *Communications of the ACM*, 50(10):94–100.

McGrath, D. K. and Gupta, M. (2008). Behind phishing: An examination of phisher modi operandi. *LEET*, 8:4.

Moghimi, M. and Varjani, A. Y. (2016). New rule-based phishing detection method. *Expert systems with applications*, 53:231–242.

Mohammad, R. M., Thabtah, F., and McCluskey, L. (2015). Tutorial and critical analysis of phishing websites methods. *Computer Science Review*, 17:1–24.

Silva, C. M. R., Feitosa, E. L., and Garcia, V. C. (2019). Heuristic-based strategy for phishing prediction: A survey of urlbased approach. *Computers & Security*.

Tsymbal, A. (2004). The problem of concept drift: definitions and related work. *Computer Science Department, Trinity College Dublin*, 106(2):58.