

Jackson José de Souza

# **Identificação de alertas de segurança virtual veiculados em redes sociais**

**São Paulo - Brasil**

**20 de outubro de 2014**



Jackson José de Souza

**Identificação de alertas de segurança virtual veiculados  
em redes sociais**

Universidade de São Paulo – USP

Instituto de Matemática e Estatística – IME-USP

Trabalho de formatura

Orientador: Daniel M. Batista

Coorientador: Elisabeti Kira

São Paulo - Brasil

20 de outubro de 2014



# Agradecimentos



# Resumo

**Palavras-chaves:** segurança computacional, redes sociais, Twitter, aprendizado de máquina.





# Lista de ilustrações

Figura 1 – Exemplo de Alerta de segurança virtual . . . . .	23
Figura 2 – Exemplo de tuíte com os tipos de metadados mais comuns no Twitter .	26
Figura 3 – Fluxograma de aprendizagem . . . . .	28



# Lista de tabelas

Tabela 1	– Dado o dicionário de palavras acima e o seguinte documento: “O modelo sacola de palavras é bem simples, pois ele lida apenas com palavras.” temos que a probabilidade da existência do documento acima dada a sacola de palavras é dada por Puni (sacola, palavras, simples, palavras) = $0.3*0.2*0.15*0.2$ . . . . .	21
Tabela 2	– Matriz de confusão . . . . .	28
Tabela 3	– Autoavaliação dos respondentes da pesquisa sobre conhecimentos de ASVs x Real conhecimento de ASVs aferido pelo controle . . . . .	52
Tabela 4	– Divisão dos respondentes que afirmaram não ter noções de segurança entre dois grupos: os que passaram no controle e os que foram reprovados no controle . . . . .	52
Tabela 5	– Divisão dos respondentes que foram reprovados no controle entre dois grupos: os que afirmaram ter noções de segurança virtual e os que afirmaram não ter noções de segurança . . . . .	52
Tabela 6	– Porcentagem média de acerto dos 10 tuítes classificados pelos respondentes dos grupos de respondentes auto declarados com ou sem noção de segurança virtual . . . . .	53
Tabela 7	– Porcentagem média de acerto dos 10 tuítes classificados pelos respondentes dos grupos de respondentes aprovados no controle e reprovados no controle . . . . .	53
Tabela 8	– Número de pessoas que acertaram a classificação do Tuíte 3 para os grupos de respondentes que declararam ter conhecimentos de segurança virtual e dos que declararam não tê-los, acompanhado de sua respectiva porcentagem em relação ao seu respectivo grupo. . . . .	53
Tabela 9	– Número de pessoas que acertaram a classificação do Tuíte 3 para os grupos de respondentes aprovados no controle e reprovados nele, acompanhado de sua respectiva porcentagem em relação ao seu respectivo grupo. . . . .	53
Tabela 10	– Número de pessoas que acertaram a classificação do Tuíte 4 para os grupos de respondentes que declararam ter conhecimentos de segurança virtual e dos que declararam não tê-los, acompanhado de sua respectiva porcentagem em relação ao seu respectivo grupo. . . . .	54
Tabela 11	– Número de pessoas que acertaram a classificação do Tuíte 4 para os grupos de respondentes aprovados no controle e reprovados nele, acompanhado de sua respectiva porcentagem em relação ao seu respectivo grupo. . . . .	54

Tabela 12 – Número de pessoas, e sua respectiva porcentagem, que acertaram a classificação dos Tuítes 1 e 5 para o grupo dos respondentes que declararam ter conhecimentos de segurança virtual e dos que declararam não tê-los. . . . .	54
Tabela 13 – Número de pessoas, e sua respectiva porcentagem, que acertaram a classificação dos Tuítes 1 e 5 para o grupo dos respondentes que foram aprovados no controle e dos que foram reprovados nele. . . . .	54
Tabela 14 – Ordem decrescente das taxas de acerto dos 2 tuítes mais fáceis e os 2 mais difíceis obtidas por cada um dos grupos de respondentes. AC: Aprovado no Controle; RC: Reprovado no Controle; CN: Com noção de segurança virtual; SN: Sem noção de segurança virtual. . . . .	55

# Sumário

<b>I</b>	<b>PARTE OBJETIVA</b>	<b>13</b>
<b>1</b>	<b>INTRODUÇÃO</b>	<b>15</b>
1.1	Motivação do trabalho	16
1.2	Objetivo	18
1.3	Estrutura do trabalho	18
<b>2</b>	<b>CONCEITOS BÁSICOS</b>	<b>19</b>
2.1	Pré-processamento e transformação de dados	19
2.2	Estatística	21
2.3	Segurança	21
2.4	Segurança virtual	22
2.5	Redes sociais e Twitter	23
2.5.1	Twitter	25
2.6	Aprendizagem Computacional	26
<b>3</b>	<b>DESENVOLVIMENTO</b>	<b>31</b>
3.1	Classes do problema	31
3.2	Coleta dos dados	31
3.3	Pré-processamento dos dados	31
3.4	Processo de classificação	32
<b>4</b>	<b>EXPERIMENTOS, DISCUSSÕES E RESULTADOS</b>	<b>33</b>
<b>5</b>	<b>CONCLUSÃO</b>	<b>35</b>
<b>II</b>	<b>PARTE SUBJETIVA</b>	<b>37</b>
	Desafios e frustrações	39
	Relação entre o trabalho e as disciplinas cursadas no BCC	41
	Trabalhos futuros	43
	Referências	45

## APÊNDICES

47

### APÊNDICE A – ANÁLISE DA PESQUISA SOBRE DETECÇÃO DE ALERTAS DE SEGURANÇA VIRTUAL DO TWITTER . . . . .

49

#### A.1 Perguntas e descrição da pesquisa . . . . .

49

##### A.1.1 População alvo e amostra . . . . .

49

##### A.1.2 Descrições das questões . . . . .

49

#### A.2 Análise e resultados da pesquisa . . . . .

51

#### A.3 Considerações finais . . . . .

56

Parte I

Parte objetiva





# 1 Introdução

O alto grau de domínio tecnológico na fabricação de computadores e a diminuição do seu custo possibilitou a popularização do seu uso pela sociedade. Essa popularização tem provocado, nas últimas décadas, uma revolução na sociedade. Tal revolução tem transformado a forma como as pessoas consomem, se comunicam, se relacionam com empresas, se entretêm etc. Como várias das atividades citadas acima necessitam da Internet para serem realizadas, a disseminação do seu uso acompanhou naturalmente o aumento do uso dos computadores.

Dessa forma, ela tornou-se um ambiente ubíquo e que pode ser acessado também por dispositivos como vídeo-games, celulares, relógios, etc. Por sua vez, as redes sociais online atraíram milhões de usuários desde a sua introdução devido ao amplo uso da Internet. Redes sociais são sites que permite interação entre os usuários cadastrados na página. As interações permitidas por uma rede social entre usuários são bastante variadas e as mais comuns são a comunicação e o compartilhamento de informações sobre assuntos de interesse em comum. Usuários podem pessoas, organizações, organizações, instituições, etc. Elas se tornaram tão populares que entre as 10 páginas mais acessadas na internet 3 são redes sociais online<sup>1</sup>. As redes sociais online serão chamadas neste trabalho de redes sociais para a leitura deste termo não cansar o leitor.

As redes sociais possuem diferentes tipos de interesses como, a comunicação entre a comunidade de uma universidade, publicidade de empresas etc. Algumas delas tentam atrair os mais variados tipos de audiência e permitem aos usuários o compartilhamento de diversos tipos de informações, desde relatos sobre o cotidiano das pessoas, fofocas sobre celebridades, até notícias importantes e de última hora. O microblog Twitter é um exemplo. Ele pode ser utilizado tanto como rede social, quanto como fonte de notícias.

O fato de de 10 bilhões<sup>2</sup> de dispositivos estarem conectados à Internet mostra a abrangência do seu uso e a sua importância como meio de comunicação. Como em todo meio de comunicação, a segurança das informações, no caso dados, transmitidas é fundamental. Afinal, apenas em um meio de comunicação seguro é possível viabilizar e assegurar a disponibilidade, a integridade, a confidencialidade e a autenticidade das informações. Porém, não há um controle rígido sobre o tráfego de dados que passa pela Internet, o que facilita que informações sejam extraídas de computadores e boa parte dos usuários não tenha ciência disso. Isso demonstra que a liberdade de comunicação proporcionada pela Internet aliada às falhas de segurança presentes nos softwares que a usam revelam um risco a segurança de pessoas, empresas, instituições etc. Este risco é grave porque as

---

<sup>1</sup> Informação obtida no dia 13/09/2013 em <<http://www.alexa.com/topsites>>

<sup>2</sup> <<http://gigaom.com/2011/10/13/internet-of-things-will-have-24-billion-devices-by-2020/>>

brechas na segurança virtual podem provocar problemas no mundo real, como grandes prejuízos financeiros às vítimas. Existem alguns softwares desenvolvidos para proteger os computadores contra falhas de segurança como anti-vírus, *firewalls*, *anti-adwares* entre outros. Alguns deles, inclusive, utilizam heurísticas para detectar ameaças que não tenham sido identificadas e catalogadas. Apesar da existência de tais softwares de segurança, a quantidade de ataques e invasões ainda causam grandes prejuízos. Estudos apontam que as perdas com crimes virtuais alcançam a casa das centenas de bilhões de dólares em prejuízos sofridos por usuários e empresas a cada ano (STRATEGIC; STUDIES, 2013).

Consequentemente, há falhas para as quais ainda não foi encontrada uma solução. Seja por não terem sido identificadas ou por serem muito recentes. Tais buracos na segurança dos softwares são críticos, pois podem ser explorados por pessoas mal intencionadas. Logo, é necessário corrigir tais brechas o quanto antes, mas para isso tais falhas devem ser identificadas. Entre as várias formas de se descobrir falhas de segurança em um software existe a identificação de alertas de segurança virtual (ASV) veiculados pela rede em sites de segurança, fóruns etc. Em (SANTOS et al., 2012) foi mostrado que é possível utilizar redes sociais para detecção de ASVs, como no próprio Twitter<sup>3</sup>. Contudo, os ASVs não são separados em uma categoria específica e não é fácil encontrá-los usando as ferramentas de busca disponibilizadas pelas redes sociais. Assim, percebeu-se que esse é um problema interessante de se abordar e é dele que este trabalho trata.

## 1.1 Motivação do trabalho

Entre o segundo trimestre de 2011 e o início de 2012 houve um aumento nas atividades hacktivistas autopromovidas no Brasil. Em maio de 2011 ocorreu um grande volume de ataques feitos pelo Anonymous a empresas privadas e agências governamentais brasileiras e em janeiro de 2012 nove grandes bancos ou agências governamentais foram atacados. O motivo do hacktivismo<sup>4</sup> no Brasil sofrer tal crescimento intenso devido ao Twitter. A disseminação do uso do Twitter no Brasil, país cuja comunidade de usuários é uma das mais assíduas<sup>5</sup>, tornou o Twitter uma boa plataforma de recrutamento para o Hacktivismo. A democratização do acesso aos computadores e à Internet, que tem sido realizada no Brasil, tornou possível que inclusive pessoas das áreas mais pobres possam ter e-mail, acesso a redes sociais, entre outros serviços disponíveis na rede. Quando o hacktivismo começou ele atraiu a atenção de muitos brasileiros. Dentre estes, vários acreditam que alguns alvos do hacktivismo mereciam sofrer os ataques e na mente dos brasileiros tais ataques não são crime. Dessa forma, o hacktivismo acabou por tocar em ponto nevrálgico, a insatisfação da população contra governos e empresas. Ao tornar

---

<sup>3</sup> <<https://www.twitter.com>>

<sup>4</sup> <https://pt.wikipedia.org/wiki/Hacktivismo> acessado em 27/07/2014

<sup>5</sup> <http://www.socialmediatoday.com/content/brazil-social-media-marketers-gold-mine>

ataques de negação de serviço (*DDoS*<sup>6</sup>) acessíveis às massas sem realizar grandes esforços, um alvo pode parar de funcionar com uma revolta popular promovida na rede. Este cenário é apenas um entre vários outros que criam o interesse de empresas e governos em monitorar redes sociais. (IMPERVA, 2012)

A necessidade de monitorar redes sociais coloca em foco outra atividade cujo objetivo é o combate e a prevenção dos ataques: a identificação de ASVs. O problema da identificação de ASVs ainda não foi explorado o suficiente e várias pessoas dentro da própria comunidade da área da computação não sabem definir adequadamente o que caracteriza um ASV. Isto foi concluído após a análise dos dados da pesquisa de identificação de ASVs do Twitter -> **colocar link para seção do apêndice**, vide apêndice, que eu realizei com o objetivo de avaliar a percepção que as pessoas possuem sobre ASVs. Alguns participantes da pesquisa manifestaram ter sentido dificuldade em fazer a classificação dos tuítes. Para exemplificar a dificuldade seguem abaixo as opiniões enviadas por 2 participantes da pesquisa:

“Muitas possibilidades para definir o que é segurança virtual. Em um universo de expressões infinitas. Virus Definitions Update Download -> Definitions Update Download is a Virus. São muito próximas as expressões, mas diferentes. A questão é o que muda nas duas?”

“É meio difícil separar o que é “alerta” mesmo (urgente, corra para se proteger/atualizar algo específico) do que é notícia relacionada com segurança (algo mais genérico, como a história dos plugins de browser), mas todos são relevantes no aspecto de segurança digital. Claro que tem que separar notícias que realmente falam de segurança daquelas que não tem nenhum conteúdo relevante nesse aspecto (ex: a da venda do exploit).”

A dificuldade de identificar alertas de segurança também se revelou nas classificações. Alguns tuítes foram classificados como alerta de segurança virtual por aproximadamente 50% dos participantes enquanto os outros cerca de 50% os classificaram como não sendo alertas de segurança virtual. Além da tarefa de identificar um ASV em publicações de redes sociais não ser simples, é inviável a um ser humano olhar cada possível alerta e classificá-lo manualmente, dado que 9.100 tuítes por segundo são postados no Twitter<sup>7</sup>. Por isso, uma solução pra o problema é treinar um sistema para identificar os alertas de segurança automaticamente.

<sup>6</sup> Um ataque de negação de serviço (também conhecido como DoS Attack, um acrônimo em inglês para Denial of Service), é uma tentativa em tornar os recursos de um sistema indisponíveis para seus utilizadores. Alvos típicos são servidores web, e o ataque tenta tornar as páginas hospedadas indisponíveis na WWW. Não se trata de uma invasão do sistema, mas sim da sua invalidação por sobrecarga. Mais detalhes em: <[https://pt.wikipedia.org/wiki/Ataque\\_de\\_nega%C3%A7%C3%A3o\\_de\\_servi%C3%A7o](https://pt.wikipedia.org/wiki/Ataque_de_nega%C3%A7%C3%A3o_de_servi%C3%A7o)>

<sup>7</sup> <<http://www.statisticbrain.com/twitter-statistics/>>

## 1.2 Objetivo

Neste trabalho é feito um estudo empírico das mensagens de segurança no Twitter escritas na língua inglesa para detectar alertas de segurança virtual. Para tal será feita uma comparação de desempenho entre os classificadores *Support vector machines* (SVM) e *Naive Bayes* para a detecção dos alertas de segurança computacional. Este estudo serve de apoio às teses de doutorado do Luiz A. F. Santos e do Rodrigo Campiolo que estão relacionadas com a detecção antecipada de anomalias em redes de computadores.

## 1.3 Estrutura do trabalho

O trabalho está dividido da seguinte forma:

- [Conceitos básicos](#)  
Introdução teórica a conceitos importantes para a compreensão do desenvolvimento do trabalho
- [Desenvolvimento](#)  
Possui a formulação do problema e o desenvolvimento teórico e prático do trabalho
- [Experimentos, discussões e resultados](#)  
Aplicação do problema, discussão do método utilizado e apresentação dos resultados
- [Conclusão](#)  
Conclusão do trabalho desenvolvido

## 2 Conceitos básicos

Este capítulo apresenta a teoria que sustenta o desenvolvimento do trabalho desenvolvido. A [seção 2.1](#) apresenta conceitos envolvendo leitura de documentos, extração de termos deles e a posterior remoção, redução, contagem de termos e transformação dos dados para a extração das características dos tuítes. A [seção 2.2](#) apresenta os conceitos de estatística necessários para compreender os métodos utilizados para a análise de texto e classificação dos tuítes. A [seção 2.3](#) define os conceitos de segurança adotados neste trabalho e a [seção 2.4](#) define os conceitos envolvendo especificamente segurança virtual para que se entenda como foram escolhidas as classes dos tuítes. A [seção 2.5](#) apresenta uma definição de rede social, explica como elas funcionam e apresenta o Twitter, suas características principais e como os usuários podem compartilhar conteúdo nele em suas várias formas. A [seção 2.6](#) introduz conceitos relacionados à aprendizagem de máquina para entender o processo de classificação dos tuítes e os resultados do processo.

### 2.1 Pré-processamento e transformação de dados

Esta seção introduz definições de filtragem e tratamento de texto incluindo definições sobre específicos conjuntos de caracteres.

Uma classe é determinada por um conjunto de elementos que possui características em comum.

URL (*Uniform Resource Locator*), ou em português Localizador-Padrão de Recursos, é a designação usada para uma cadeia (conjunto) de caracteres que indicam a localização de um recurso em uma rede. Neste trabalho a URL pode ser entendida como o endereço de uma página na Internet. URL curta (*short URL*) é um endereço reduzido de uma página que costuma ser utilizado para referenciar o endereço original de um site em textos cujo limite de caracteres para escrita é reduzido como o Twitter. Quando se deseja obter uma URL curta pode-se usar um serviço de encurtamento de URLs como o <http://tinyurl.com/>.

Tokenização é o ato de decompor um documento em peças chamadas tokens. As peças são ocorrências específicas de cadeias de caracteres e são separadas por caracteres chamados delimitadores. Eis um exemplo:

Delimitadores (entre aspas simples): ‘,’ ‘;’ ‘ ’

Documento: Friends, Romans, Countrymen, lend me your ears;

Tokens:

Note que o caractere espaço ‘ ’ é também um delimitador de token.

Friends	Romans	Countrymen	lend	me	your	ears
---------	--------	------------	------	----	------	------

Normalização de tokens é a tarefa de converter em uma forma canônica tokens que sejam diferentes mas que possuam um mesmo significado. Assim, temos o token canônico que dá nome à sua classe de equivalência e quaisquer tokens que sejam convertíveis ao canônico pertencem à mesma classe de equivalência. Qualquer token que pertence a uma dada classe de equivalência é representante dela e deve casar com qualquer outro token da classe, embora haja diferenças entre as cadeias de caracteres dos seus tokens.

Exemplos:

naive => Naive, Naïve, naïve, NAÏVE, NAIVE

usa => U.S.A., USA

Radicalizamento é o processo de reduzir palavras flexionadas ou derivadas à uma forma básica comum, o radical. O radical neste caso não precisa ser igual ao seu homônimo linguístico.

Exemplo: real, reais, realizar, realizável, realista => rea

Metadado é um dado sobre um dado. De outra forma, podemos dizer que um metadado é uma informação sobre um dado. Por exemplo, quando se busca tuítes podemos utilizar filtros que exibam como resposta apenas os tuítes em língua inglesa ou que tenham sido escritos no dia 17/04/2013. A língua e a data são exemplos de metadados que um tuíte possui, pois são informações sobre um tuíte que é o documento (dado) de interesse.

Aqui consideraremos como tipo uma classe de todos os tokens que possuem exatamente a mesma cadeia de caracteres e que faz parte do dicionário de classificação. Em outras palavras, um token é uma cópia de um tipo de forma que um documento pode conter várias cópias do seu representante, mas um tipo é único.

Frequência de um termo é o número de vezes que um termo ocorre em um documento e é denotado por  $tft,d$  (transformar em fórmula).

O modelo sacola de palavras é uma representação simplificadora utilizada em problemas de classificação. No modelo o documento ou objeto é representado por uma coleção de termos (ou palavras) sem preocupação com a ordem dos termos. Porém, o número de ocorrência de cada termo é guardada.

Modelo de língua unigrama considera que a probabilidade da ocorrência de cada termo é independente da ocorrência prévia de quaisquer termos. Ou seja, a ordem da ocorrência dos termos não importa. Então, temos:  $Puni(t_1 \dots t_n) = \text{produtório da probabilidade de cada termo}$

OBS: Vale ressaltar que os vários conceitos de tratamento mencionados acima podem ser aplicados de várias formas e isso depende do programa de computador utilizado.

Termo	Probabilidade
sacola	0.3
palavras	0.2
simples	0.15

Tabela 1 – Dado o dicionário de palavras acima e o seguinte documento: “O modelo sacola de palavras é bem simples, pois ele lida apenas com palavras.” temos que a probabilidade da existência do documento acima dada a sacola de palavras é dada por Puni (sacola, palavras, simples, palavras) =  $0.3 \cdot 0.2 \cdot 0.15 \cdot 0.2$

Ou seja, há mais de uma forma de se fazer radicalizamento, o que depende dos objetivos da necessidade do conteúdo do texto que está sendo analisado.

## 2.2 Estatística

Quando temos  $n$  eventos que são independentes entre si a probabilidade deles ocorrerem simultaneamente é dada por:

$$P(E_1 \cap E_2 \cap \dots E_n) = \prod_{i=1}^n P(E_i) = P(E_1) * P(E_2) * \dots P(E_n)$$

## 2.3 Segurança

Segurança consiste basicamente na proteção de um bem seja ele material ou imaterial e compreende também o controle de ameaças a tal bem. Um bem material pode ser uma pessoa, uma casa e um bem imaterial seria o conhecimento ou a cultura de um povo, por exemplo.

A preocupação das sociedades, instituições e países em proteger vários tipos de bens inspirou a divisão de tais bens em diversas categorias.

- **Segurança do interior**

Esta categoria, também chamada em inglês de *Homeland security*, se refere aos esforços nacionais para prevenir ataques terroristas, reduzir a vulnerabilidade de um país ao terrorismo e minimizar os danos consequentes de ataques e desastres naturais que ocorrerem. Esta categoria foi adaptada à preocupação atual dos EUA com o risco de ataques terroristas em seu país.

- **Segurança pública**

“A Segurança Pública é uma atividade pertinente aos órgãos estatais e à comunidade como um todo, realizada com o fito de proteger a cidadania, prevenindo e controlando manifestações da criminalidade e da violência, efetivas ou potenciais, garantindo o exercício pleno da cidadania nos limites da lei.” (MINISTÉRIO... , 2013).

- **Segurança nacional**

Trata-se do estado mensurável da capacidade de uma nação superar as múltiplas ameaças ao aparente bem estar da sua população e sua sobrevivência como um Estado-nação, a qualquer momento. Isto se faz pelo balanceamento da política de estado através da governança, que pode ser guiada pela computação, empiricamente ou de outra forma, e é extensível à segurança global por variáveis externas ao governo ([PALERI, 2008](#)).

- **Segurança física**

Segurança física compreende as medidas adotadas para negar acesso não autorizado a instalações, equipamentos e recursos, e proteger o pessoal e propriedade contra perdas e danos provocados por espionagem, roubo, ataques terroristas e desastres naturais. Traduzido e adaptado de ([HEADQUARTERS, 2001](#)).

Além das categorias citadas acima também existe o que se decidiu chamar de segurança virtual e é a esta categoria de segurança que é dedicada à [seção 2.4](#).

## 2.4 Segurança virtual

Esta seção aborda alguns conceitos e possui definições envolvendo segurança virtual que serão utilizados na classificação dos tuítes. Alguns dos conceitos e definições desta seção são baseados em ([FUTURE, 2011](#); [SECURITY, 2010](#); [WILSHUSEN, 2013](#); [CERT.PT; CSIRT, 2012](#); [SHIRLEY, 2007](#); [WILSHUSEN, 2011](#); [GLOSSARY... , 2009](#)).

Segurança virtual consiste na prevenção de dano, proteção contra uso não autorizado, exploração e também envolve a restauração de dados, sistemas de comunicação e de informação para garantir confidencialidade, integridade e acessibilidade de dados e digitais programas de computador.

Um incidente de segurança virtual (ISV) pode ser considerado como um evento adverso, confirmado ou sob suspeita, que tem por consequência o acesso, extração, manipulação ou corrompimento da integridade, confidencialidade, segurança ou acessibilidade de dados ou programas de computador, sejam públicos ou privados, sem autorização legal. Um evento pode ser causado intencional ou não intencionalmente, ter um alvo específico ou vago, e pode fazer uso de variadas técnicas. Ele pode surgir a partir de diferentes fontes, incluindo um país fazendo espionagem ou guerra de informações contra outros países, criminosos, *crackers*, programadores de vírus, terroristas entre outros.

ISVs não intencionais podem ser causados por erro ou omissão humana e falhas de equipamentos, como por exemplo, a operação de um sistema por funcionários displicentes ou sem treinamento, atualizações de programa de computador, realização de manutenções entre outros. ISVs não intencionais podem corromper dados ou provocar interrupções ou



mau funcionamento de sistemas. ISVs intencionais são provocados por um ente inteligente, como um *cracker* ou organização criminosa, e incluem ataques com alvo específico ou vago. Um ataque com alvo específico ocorre quando um grupo de pessoas ou um único indivíduo realiza um ataque contra um sistema de infraestrutura crítica. Um ataque de alvo vago ocorre quando o alvo definido para a realização do ataque não é, a princípio, claro como é o caso de um vírus, *worm* ou *malware* liberado na internet sem alvo específico.

No contexto de segurança virtual um ataque consiste na tentativa de destruir, expor, alterar ou incapacitar algum software, sistema e ou dados contidos neles, ou qualquer outra falha de segurança em dispositivos eletrônicos (GLOSSARY..., 2009 apud STANDARDIZATION, 2006).

Há algumas formas de estruturar a classificação dos eventos e incidentes como em (WILSHUSEN, 2013; CERT.PT; CSIRT, 2012). Neste trabalho vamos adotar a classificação de ISV utilizada em (CERT.PT; CSIRT, 2012). Existem várias classes e tipos de incidentes que agrupam tipos de eventos. Para conhecer mais os tipos de eventos veja (CERT.PT; CSIRT, 2012).

Define-se um alerta de segurança virtual (ASV) como um aviso, geralmente de caráter urgente, sobre a ameaça, ocorrência, uma notícia de solução para, uso de ferramenta para, ou a explicação de como gerar um ISV.

Um exemplo de ASV pode ser visto no tuíte<sup>1</sup> a seguir:



Figura 1 – Exemplo de Alerta de segurança virtual

<sup>1</sup> Todos os tuítes deste trabalho seguem as regras de publicação de tuítes em trabalhos segundo o Twitter. Ver <<https://twitter.com/logo>> seção: Offline (static uses and publications)

## 2.5 Redes sociais e Twitter

Esta seção faz uma introdução ao tema das redes sociais e coloca em destaque o Twitter, que é a fonte dos dados utilizados no trabalho. Assim, esta seção explica o que é um tuíte, qual o seu conteúdo, e como eles se propagam dentro do Twitter. As fontes utilizadas na escrita desta seção foram (BOYD; ELLISON, 2007; TWITTER. . . , 2013).

Redes sociais são serviços hospedados na Internet que permitem a indivíduos construir um perfil, se expressar a outros indivíduos com os quais ele possui alguma conexão na rede social, visualizar sua lista de conexões e participar de outras listas que tenham sido criadas por outros. As conexões podem ser os amigos em uma rede social e as listas podem ser conjuntos de integrantes de um grupo, de evento etc. Vale mencionar que a classificação de um relacionamento em uma rede social como sendo o de amizade não significa que os usuários realmente sejam amigos no sentido denotativo da palavra.

A visibilidade do conteúdo gerado ou compartilhado por usuários depende das restrições impostas pela rede social e pelos usuários. Por exemplo, o perfil do usuário pode ser total ou parcialmente público e a sua visibilidade na Internet depende de como o provedor do serviço controla as informações da rede social. Em outras palavras, o perfil do usuário pode não ser visível a todos na Internet, ou seja, aos indivíduos que não são usuários da rede social na qual o perfil foi criado. O mesmo acontece com outros tipos de conteúdo publicados na rede social. O conteúdo pode ser público para os integrantes da rede social, mas invisível a não usuários dela, também como pode ser restrito a determinadas conexões ou a todos os usuários que não possuem conexão com o divulgador do conteúdo.

As conexões podem ser estabelecidas de forma unidirecional e bidirecional. Ou seja, uma conexão bidirecional depende da aprovação de ambos os usuários envolvidos nela a respeito do status do relacionamento enquanto que a unidirecional depende apenas da vontade de um usuário. Em algumas redes sociais para duas pessoas serem amigas é necessário o consentimento de ambas, mas um usuário pode liberar o acesso do conteúdo que ele publica a outros usuários sem que estes façam o mesmo.

As comunicações entre conexões podem ser feitas de várias formas. Entre elas, existe a troca de mensagens visível a todos os usuários da rede, aos usuários conectados a pelo menos um dos participantes da troca de mensagens e apenas entre os participantes da troca de mensagens. O tipo de conteúdo publicado varia desde texto em língua natural até foto, áudio, vídeo entre outros tipos de conteúdo. Finalmente, como as redes sociais possuem o intuito de serem o mais acessíveis possível, elas possibilitam o seu uso por meio de computadores de mesa, notebooks, *smartphones* e até aparelhos celulares comuns.

O que torna as redes sociais únicas é o fato de que elas não apenas permitem que indivíduos conheçam estranhos, mas também tornam possível aos usuários se comunica-

rem e tornar visíveis seus grupos de conexões. Ao se expressar dentro da rede social o conteúdo gerado pelo usuário pode se dispersar entre todos os seus grupos de conexões e isto possibilita a criação de conexões entre indivíduos, que não seriam possíveis de outra forma. Apesar de, em geral, não existir o objetivo de se criar tais conexões a publicação de conteúdo na rede social faz com que as conversas travadas com os outros usuários da rede social tenham como consequência a criação de novas conexões na rede social devido à existência de interesses em comum entre indivíduos que não se conhecem pessoalmente.

Porém, vale ressaltar que os usuários, em várias das redes sociais, não estão necessariamente buscando fazer troca de conhecimento com pessoas que possuem interesses em comum nem buscando criar novas conexões. Na verdade, elas estão basicamente se comunicando com as suas redes de conexões, ou contatos.

### 2.5.1 Twitter

O Twitter é uma rede social que funciona também como microblog permitindo aos usuários lerem e enviarem tuítes. Tuítes são mensagens de texto com até 140 caracteres. Os tuítes podem ser enviados por meio de aplicativos para *smartphones*, página na Internet ou por SMS, em alguns países. Os tuítes dos usuários, por padrão, são visíveis a qualquer um que tenha acesso a Internet, mas seu acesso também pode ser limitado apenas aos usuários conectados ao usuário que envia tais mensagens, os seguidores. Um seguidor no Twitter é uma conexão que possui permissão para ler as mensagens de um dado usuário e permite a esse usuário enviar ‘mensagens diretas’ ao seguidor. Se um usuário quiser ele pode deixar de seguir alguém ou bloquear algum seguidor. ‘Mensagens diretas’ são tuítes que apenas o remetente e o destinatário podem ver. Usuários podem ser pessoas, empresas, instituições etc.

Os tuítes podem ser agrupados por assunto utilizando *hashtags* - palavras ou frases precedidas de uma cerquilha ‘#’. Da mesma forma, é possível mencionar um usuário em um tuíte usando um ‘@’ sucedido do nome de um usuário (sem espaços). Isto significa que o usuário mencionado poderá ver tal tuíte. Uma resposta é um caso particular de menção em que o tuíte começa com o ‘@’ seguido do nome do usuário ao qual se está respondendo. Além disso, também é possível replicar um tuíte contanto que o seu dono não tenha limitado seu acesso apenas a seus seguidores. Os tuítes replicados possuem o acrônimo RT (*retweet*) seguido pela menção do usuário que originalmente escreveu o tuíte. Outra característica comum nos tuítes é o uso de URLs curtas devido ao limite do número de caracteres de um tuíte. Caso o usuário esteja escrevendo um tuíte com uma URL não encurtada (mais de 20 caracteres) o Twitter utiliza seu próprio encurtador. Assim, o tamanho do tuíte é contado de acordo com o número de caracteres da URL encurtada em substituição à original.

Eis um exemplo de um tuíte que reúne todos os tipos de metadados supracitados:



Figura 2 – Exemplo de tuíte com os tipos de metadados mais comuns no Twitter

Além do tuíte clássico de 140 caracteres também é possível publicar tuítes expandidos. Estes tuítes podem conter fotos, vídeos e cartões. Estes conteúdos multimídia podem ser adicionados usando a própria plataforma do Twitter, para fazer o *upload* de fotos, ou aplicativos como o [<https://vine.co/>](https://vine.co/) para inserir vídeos no tuíte. O conteúdo multimídia é disponibilizado via um link, respeitando o limite de caracteres de um tuíte. Um cartão de tuíte ou *tweet card* é um conteúdo multimídia expandido utilizado para exibir fotos, vídeos, propaganda, resumo de notícias etc. O cartão é gerado a partir da inserção de alguns metadados no tuíte que permitem visualizar o conteúdo. Note que este é um meio adicional utilizado para publicar fotos e vídeos no Twitter, mas não é o único.

## 2.6 Aprendizagem Computacional

Aprendizagem de Máquina é uma subárea de Inteligência Artificial cujo objetivo é construir sistemas que são treinados a executar uma dada tarefa aperfeiçoando o seu desempenho conforme ganham experiência em realizar tal tarefa. Aprendizagem de Máquina é uma técnica em que se busca encontrar padrões nos dados relacionados à tarefa de interesse e são definidas regras ou maneiras de utilizar tais dados extraídos para que o sistema possa executar a tarefa de forma satisfatória. O aprendizado é bem sucedido se ele se aperfeiçoa conforme aumenta a exposição aos dados relacionados ao problema que deve ser resolvido.

Por exemplo, um sistema que utiliza Aprendizagem de Máquina pode ser treinado para distinguir mensagens enviadas por e-mail e separá-las em mensagens spam e não spam. Conforme o sistema possui mais mensagens spam ou não-spam (dados) ele consegue separar de forma cada vez mais próxima do ideal as mensagens spam das não-spam. Após

os resultados do treinamento nesta tarefa serem considerados satisfatórios o sistema será utilizado para separar, ou classificar, as mensagens de e-mail recebidas por um usuário que irão para a caixa de entrada e as que irão para a pasta de spam.

Também é possível definir a aprendizagem de máquina de forma mais geral: DANIEL: quero deixar como está abaixo, mas vou tirar anotação de tabela =p

*Um programa de computador aprende a partir da experiência  $E$  com respeito a uma classe de tarefas  $T$  e medida de desempenho  $P$ , se a sua performance em  $T$ , segundo a medição  $P$ , melhora com a experiência  $E$ .*

Tradução livre de (MITCHELL, 1997)

Em aprendizagem há algumas formas de aprendizado. Para efeitos de simplificação divide-se grosseiramente a aprendizagem em supervisionada e não-supervisionada. Na aprendizagem supervisionada o sistema recebe um conjunto de dados de treinamento e outro de testes, ambos separados por categorias, as classes. O conjunto de treinamento é utilizado para aprender a identificar as categorias dos dados e o conjunto de testes é utilizado pra testar se as categorias dos dados são identificadas corretamente. A detecção de spam é um caso de aprendizado supervisionado.

O aprendizado não-supervisionado recebe um conjunto de dados não categorizados e tenta separar os dados em grupos cujos dados compartilham características próprias do grupo. Como as categorias dos dados não são conhecidas é difícil avaliar o desempenho do sistema na separação dos dados.

Classificação: Dado um conjunto de classes busca-se determinar a qual classe um objeto pertence. Um problema de classificação possui 2 ou mais classes e embora na maioria dos problemas um objeto pertença a apenas um classe também é possível atribuir mais de uma classe a cada objeto. Um modelo de aprendizagem de máquina que faz a classificação de objetos é chamado de classificador.

Categorização é o ato de atribuir uma classe a cada objeto utilizado na construção do classificador, o que envolve o conjunto de treinamento, teste e validação do classificador.

Daniel, a definição dos itens abaixo está incompleta.

Um problema de classificação possui 6 fases:

- A aquisição dos dados consiste na coleta e categorização dos dados.
- Pré-processamento é a fase em que são removidos os
- Extração das características
- Seleção do modelo de aprendizagem

- Treinamento do classificador é fase em o modelo adotado irá utilizar o conjunto de treinamento para aprender os critérios de decisão irá utilizar para realizar a classificação.
- Ajuste das características e validação do classificador

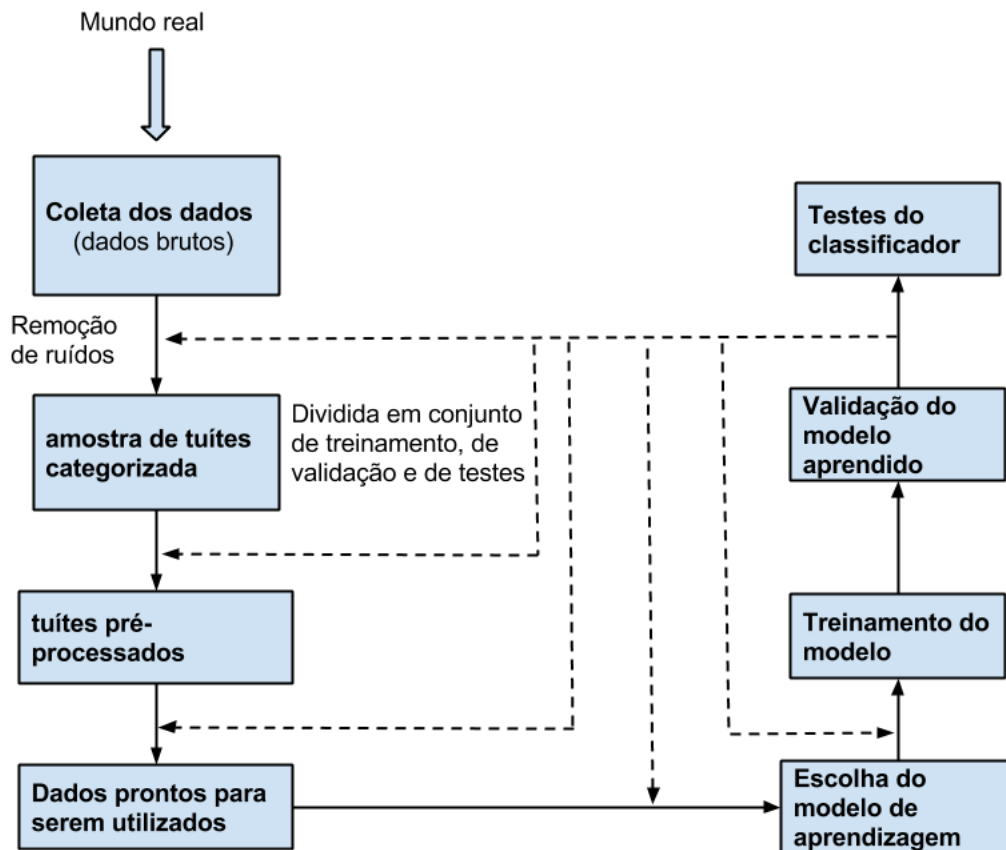


Figura 3 – Fluxograma de aprendizagem

Tabela 2 – Matriz de confusão

		Classe inferida	
		Sim	Não
Classe do tuíte	Sim	verdadeiro positivo - $vp$	falso positivo - $fp$
	Não	falso negativo - $fn$	verdadeiro negativo - $vn$

Precisão:

$$\frac{pv}{pv + fp}$$

Por exemplo de documentos classificados como pertencentes à classe  $x$  e que pertencem a  $x$  / número de documentos que foram classificados como pertencentes a uma classe  $x$ .

Recuperação (*Recall*):

$$\frac{pv}{pv + fn}$$

Por exemplo de documentos classificados como pertencentes à classe  $x$  e que pertencem a  $x$  / de documentos pertencentes à classe  $x$

Acurácia:

$$\frac{pv + nv}{pv + fp + fn + nv}$$

Por exemplo: de documentos classificados como pertencentes à classe  $x$  e que pertencem a  $x$  + de documentos classificados como não pertencentes à classe  $x$  e que não pertencem a  $x$  / todos os documentos Em outras palavras é o número de documentos classificados corretamente





## 3 Desenvolvimento

Explicar sobre o porque dos tuítes serem em inglês

### 3.1 Classes do problema

definição das classes

### 3.2 Coleta dos dados

Número de tuítes separados por classes

explicar como foi feita o etiquetamento dos tuítes

Os tuítes foram coletados em duas fases: A 1ª base de dados possui 260440 e foi coletada usando blá entre 28/04/2012 a 10/02/2013 e posteriormente foram coletados também usando blá 40307 entre 14/06/2013 a 31/07/2013.

As duas bases de dados, agrupadas em um arquivo json, foram unidas em uma única base de dados da qual são extraídos os tuítes do estudo.

### 3.3 Pré-processamento dos dados

Para pré-processar os dados é necessário conhecê-los. Ou seja, como são representados, o significado dos metadados presentes neles e os tipos de ruído que eles possuem. É importante destacar que o filtro da api do Twitter utilizado para coletar os tuítes possui uma detecção que não funciona muito bem para identificar o idioma em que o tuíte foi escrito. Por isso foi utilizado um classificador que identifica o idioma do tuíte.

Levando em conta as ferramentas disponíveis para efetuar a filtragem dos dados o pré-processamento dos dados foi feito em duas fases:

1. Um script ruby olha o texto de cada tuíte, pega o texto presente na headline de cada página cujo link se encontra no texto do tuíte e adiciona ao tuíte. Depois são decodificados os caracteres html presentes no texto do tuíte e após isso são removidos os tuítes que não possuem informações suficientes para serem classificados segundo os critérios definidos abaixo.
2. Script que processa os tuítes e gera um arquivo

explicar o tipo de processamento que é feito nos textos dos tuítes

Critérios que um tuíte deve satisfazer para ser considerado no problema:

### 3.4 Processo de classificação

Apresentar a Weka

Apresentar os classificadores -> Naive Bayes (multinomial) e SVM

Uso do modelo bag of Words

Como a ordem dos termos na sacola de palavras não importa, então qualquer ordenação de um conjunto de termos possui a mesma probabilidade de ocorrência. Assim, temos que a probabilidade de ocorrência de tais termos segue a distribuição multinomial.

## 4 Experimentos, discussões e resultados

Descrição do uso da Weka para executar o algoritmo de aprendizagem...

Mostrar algumas estatísticas úteis sobre os tuítes

Número de atributos

Uso de Information Gain para melhorar a classificação

Naive Bayes Multinomial

Tuítes corretamente classificados	2382	76.7892%
Tuítes incorretamente classificados	720	23.2108%

Taxa VP	Taxa FP	Precision	Recall	F-Measure	Classe
0.662	0.07	0.562	0.662	0.608	Notícia de segurança virtual
0.688	0.025	0.762	0.688	0.723	Alerta de seg. virt. e seg. não virt.
0.793	0.107	0.781	0.793	0.787	Alerta de segurança virtual
0.536	0.007	0.712	0.536	0.612	Notícia de seg. geral e seg. virtual
0.949	0.043	0.907	0.949	0.927	Notícia de segurança geral
0.495	0.018	0.639	0.495	0.558	Potencial alerta de seg. virtual
0.407	0.024	0.493	0.407	0.446	Spam
0.768	0.061	0.765	0.768	0.765	Média ponderada das classes

a	b	c	d	e	f	g	
245	4	78	1	12	13	17	a = Notícia de segurança virtual
15	221	54	5	16	9	1	b = Alerta de seg. virt. e seg. não virt.
74	48	797	6	25	27	28	c = Alerta de segurança virtual
12	7	7	52	16	0	3	d = Notícia de seg. geral e seg. virtual
6	4	13	7	907	0	19	e = Notícia de segurança geral
25	4	59	0	4	92	2	f = Potencial alerta de seg. virtual
59	2	13	2	20	3	68	g = Spam

Matriz de confusão

SVM

Tuítes corretamente classificados	2338	75.3707 %
Tuítes incorretamente classificados	764	24.6293 %

Taxa VP	Taxa FP	Precision	Recall	F-Measure	Classe
0.595	0.075	0.519	0.595	0.554	Notícia de segurança virtual
0.648	0.02	0.788	0.648	0.711	Alerta de seg. virt. e seg. não virt.
0.803	0.134	0.742	0.803	0.772	Alerta de segurança virtual
0.567	0.007	0.733	0.567	0.64	Notícia de seg. geral e seg. virtual
0.932	0.041	0.911	0.932	0.921	Notícia de segurança geral
0.478	0.023	0.574	0.478	0.522	Potencial alerta de seg. virtual
0.407	0.017	0.571	0.407	0.476	Spam
0.754	0.069	0.753	0.754	0.751	Média ponderada das classes

a	b	c	d	e	f	g	
220	4	95	3	13	15	20	a = Notícia de segurança virtual
18	208	63	4	15	12	1	b = Alerta de seg. virt. e seg. não virt.
88	33	807	6	24	34	13	c = Alerta de segurança virtual
12	9	14	55	6	1	0	d = Notícia de seg. geral e seg. virtual
15	4	25	4	891	1	16	e = Notícia de segurança geral
29	3	58	1	5	89	1	f = Potencial alerta de seg. virtual
42	3	25	2	24	3	68	g = Spam

Matriz de confusão

## 5 Conclusão



## Parte II

### Parte subjetiva





# Desafios e frustrações

Demora para entender como processar o texto corretamente

Falta de conhecimento em segurança computacional -> foram jogados fora 2 mil tuítes por duas vezes por considerar que minha compreensão do que é um alerta de segurança estava imatura

\*\*\*dificuldade de me dedicar ao projeto o quanto gostaria devido às outras atividades que eu conduzi durante o trabalho\*\*\* -> esquecimento de partes do trabalho, o que me levou a ter que reler bibliografias e rever o que estava sendo feito ao longo do projeto

Desvio de foco também me atrapalhou -> gastar bastante tempo com um classificador de línguas que não funcionou como esperado ao invés de cuidar mais do problema real



## Relação entre o trabalho e as disciplinas cursadas no BCC

MAC0460 Aprendizagem Computacional: Modelos, Algoritmos e Aplicações

MAC0110 Introdução à Computação

MAC0122 Princípios de Desenvolvimento de Algoritmos

MAC0211 Laboratório de Programação I

MAE0121 Introdução à Probabilidade e à Estatística I

MAE0212 Introdução à Probabilidade e à Estatística II

MAT0111 Cálculo Diferencial e Integral I

MAC0459 Ciência e Engenharia de Dados



## Trabalhos futuros

Verificar se os próprios tuítes são uma ameaça de segurança. Por exemplo, tuítes do Blackhole exploit kit -> referência ao artigo da Jeanna

Capturar o título das páginas de links presentes nos tuítes para melhorar o desempenho do classificador

Usar o Open Calais para identificar entidades presentes nos tuítes a fim de melhorar o classificador



# Referências

BOYD, D. M.; ELLISON, N. B. Social network sites: Definition, history, and scholarship. *J. Computer-Mediated Communication*, v. 13, n. 1, p. 210–230, 2007. Citado na página 24.

CERT.PT; CSIRT, R. N. *Taxonomia Comum para a Rede Nacional de CSIRTs*. 2012. <[www.cert.pt/images/docs/Taxonomiav2.5.pdf](http://www.cert.pt/images/docs/Taxonomiav2.5.pdf)>. Citado 2 vezes nas páginas 22 e 23.

FUTURE, D. of H. S. *Blueprint for a Secure Cyber Future: The Cybersecurity Strategy for the Homeland Security Enterprise*. [S.l.], 2011. Citado na página 22.

GLOSSARY of IT Security Terminology Terms and definitions. [S.l.]: TeleTrust Germany, 2009. <[http://www.teletrust.de/uploads/media/ISOIEC\\_JTC1\\_SC27\\_IT\\_Security\\_Glossary\\_TeleTrust\\_Documentation.pdf](http://www.teletrust.de/uploads/media/ISOIEC_JTC1_SC27_IT_Security_Glossary_TeleTrust_Documentation.pdf)>. Citado 2 vezes nas páginas 22 e 23.

HEADQUARTERS, U. S. D. of A. *Field Manual 3-19.30: Physical Security*. 2001. <<http://www.globalsecurity.org/military/library/policy/army/fm/3-19-30/ch1.htm>>. Capítulo 1. Citado na página 22.

IMPERVA. *Imperva's Hacker Intelligence Summary Report: The Anatomy of an Anonymous Attack*. [S.l.], 2012. Citado na página 17.

MINISTÉRIO da Justiça - Órgão de Segurança. 2013. <<http://portal.mj.gov.br/>> em Órgãos de Segurança, Conceitos Básicos. Acessado: 26-10-13. Citado na página 21.

MITCHELL, T. M. *Machine Learning*. 1. ed. New York, NY, USA: McGraw-Hill, Inc., 1997. ISBN 0070428077, 9780070428072. Citado na página 27.

PALERI, P. National security: Imperatives and challenges. In: \_\_\_\_\_. [S.l.]: Tata McGraw-Hill, 2008. p. 57. ISBN 9780070656864. Citado na página 22.

SANTOS, L. A. F. et al. Análise de mensagens de segurança postadas no twitter. *Anais do simpósio brasileiro de sistemas colaborativos (SBSC)*, n. 3, p. 20–28, 2012. Citado na página 16.

SECURITY, U. D. of H. *Privacy Impact Assessment for the Initiative Three Exercise*. [S.l.], 2010. Citado na página 22.

SHIRLEY, R. *Internet Security Glossary*. 2007. <<http://www.ipa.go.jp/security/rfc/RFC4949-00EN.html>>. Citado na página 22.

STRATEGIC, C. for; STUDIES, I. *The Economic Impact of Cybercrime and Cyber Espionage*. [S.l.], 2013. Citado na página 16.

TWITTER Help Center - Get started: FAQs and the basics. 2013. <<https://support.twitter.com/groups/50-welcome-to-twitter>>. Acessado: 28-09-13. Citado na página 24.

WILSHUSEN, G. C. *Cybersecurity: Continued Attention Needed to Protect Our Nation's Critical Infrastructure*. [S.l.], 2011. Citado na página [22](#).

WILSHUSEN, G. C. *Cyber Threats Facilitate Ability to Commit Economic Espionage*. [S.l.], 2013. Citado 2 vezes nas páginas [22](#) e [23](#).



## Apêndices



# APÊNDICE A – Análise da pesquisa sobre detecção de Alertas de segurança virtual do Twitter

## A.1 Perguntas e descrição da pesquisa

A pesquisa consiste de 13 questões. Entre elas há 10 tuítes e as outras 3 questões servem para filtrar os elementos da amostra (participantes que estudam computação ou trabalham na área), saber se o mesmo considera ter conhecimentos sobre ASVs e coletar sugestões e críticas sobre a pesquisa. Também é associado a cada participante da pesquisa o tempo, em segundos, gasto para preencher a pesquisa.

Todas as perguntas da pesquisa são obrigatórias exceto a que solicita o envio de sugestões ou críticas.

### A.1.1 População alvo e amostra

A população de interesse são alunos de computação, professores e profissionais da área de computação sem necessariamente possuírem experiência em segurança virtual (*cybersecurity*). Para se obter a amostra, foram contactados os alunos e professores do IME e profissionais fora da comunidade USP via e-mail e redes sociais e, por sua vez, algumas das pessoas contactadas enviaram a pesquisa a conhecidos e colegas relacionados à área de computação.

Assim, observa-se que a amostra não é aleatória no seu sentido literal, pois ela é composta pelas pessoas com as quais eu consegui entrar em contato e algumas outras que souberam da pesquisa por meio de alguém que já tinha sido contactado por mim. Todos os respondentes participaram da pesquisa voluntariamente.

### A.1.2 Descrições das questões

Os 10 tuítes utilizados na pesquisa estão todos escritos em língua inglesa. Abaixo a lista de tuítes e as respostas esperadas na classificação deles:

1. “How secure are Apple’s iPhone and iPad from malware, really? | Naked Security <http://t.co/rlyCYs7H>” – ASV

2. “The heart attack I get when the security alarm sensor falls and I think / someones coming in...had my...” – não-ASV
3. “RT @securityaffairs: DDoS attacks in Q2, do not underestimate the cyber threat <http://t.co/04ThaJ0q...>” – ASV
4. “Researchers find malware targeting online stock trading software: Security research... <http://t.co...>” – ASV
5. “#Apple iOS Zero-day exploit sold for \$500,000 <http://t.co/NOFfIM4D36> #0day #iphone #security #vulner...” – ASV
6. “Possibilities for Malicious Browser Extensions Are Almost Infinite, Researcher Says: The Hacker Halt...” – ASV
7. “Lubbock, Tech officials mainstening security after Boston attack <http://t.co/dIRfj9UbGC>” – não-ASV
8. “Sen. Rockefeller questions cyber security of critical infrastructure after attack on gas pipelines:...” – não-ASV
9. “Homeland Security will track this article if I say electric pork cloud virus. oops. <http://t.co/AcCK...>” – não-ASV
10. “Avast! Avast Antivirus 8.0.1489 Virus Definitions Update Download (August 24, 2013) KEYGURU <http://n...>” – não-ASV

As classes às quais os tuítes podem pertencer são mutuamente exclusivas e elas são as seguintes: ASVs, Spam, notícia de segurança geral e notícia de segurança virtual. Porém, para todos os efeitos, nesta pesquisa as classes adotadas são ASV e não-ASV, que é formada pelos grupos restantes de tuítes. Os tuítes foram divididos entre 5 ASVs e 5 não-ASVs.

Os participantes classificam cada um dos tuítes utilizando a noção que cada um deles possui sobre ASVs. Ou seja, as classificações são feitas sem que lhes tenha sido apresentada uma definição de ASVs.

As respostas possíveis para cada tuíte são:

1. Alerta de segurança virtual
2. Não é alerta de segurança virtual
3. Não sei

A opção “Não sei” pode incluir um comentário opcional do respondente.

Antes de fazer a classificação de cada tuíte são apresentadas duas questões sobre o perfil do respondente, além da inclusão do tempo gasto pelo participante para preencher a pesquisa. O tempo associado a cada participante é registrado pela Qualtrics<sup>1</sup>, plataforma que armazena as respostas da pesquisa. A primeira pergunta solicita a área de atuação do respondente, pois apenas pessoas relacionadas à área de computação (estudantes de qualquer nível educacional, profissionais ou docentes/pesquisadores) serão analisadas na amostra coletada por se considerar que tais pessoas são, em geral, mais capacitadas para fazer a classificação dos tuítes (possíveis ASVs). A segunda pergunta busca verificar se o respondente considera ter (ou não) conhecimentos sobre ASVs. O tempo de preenchimento da pesquisa é utilizado para aferir se o participante teve tempo suficiente para ler e pensar na escolha das classificações dos tuítes. O tempo passa a ser contado a partir do momento em que o participante interage com a pesquisa passando o cursor na aba em que ela foi aberta ou digitando alguma tecla. A sua contagem é finalizada quando as respostas são enviadas à plataforma da pesquisa.

## A.2 Análise e resultados da pesquisa

A análise dos resultados foi feita utilizando questões de controle e as respostas da questão em que o respondente declara se possui conhecimento sobre segurança virtual. Para fins de identificação de ASVs e não-ASVs considera-se satisfatória uma taxa de acerto superior a 80%

Foram escolhidos para servir de controle os Tuítes 2 e 7, que devem ser classificados como não-ASV. O objetivo do controle é separar os participantes que conseguem responder ambos os tuítes corretamente dos que não conseguem. Os respondentes reprovados no controle são considerados como possuidores de um baixo nível de familiaridade com ameaças de segurança virtual e suas respostas devem ser observadas com cuidado.

Ao todo 100 respondentes preencheram a pesquisa. Para compor a amostra a ser analisada foram removidas as respostas de 4 respondentes porque tratam-se de participantes que não possuem atuação na área da computação ou correlatas. Após a realização da análise da amostra de 96 pessoas apenas 4 (4,16%) delas, dentre as quais todas afirmaram ter noções de segurança virtual, acertaram todas as classificações de tuítes.

Os cinco participantes que preencheram a pesquisa muito rápido (menos que 60 segundos) provavelmente não refletiram para responder algumas questões e escolheram suas respostas de maneira aleatória ou não consciente. Os três participantes que levaram muito tempo (mais de 1h30) pra responder o questionário podem ter começado a preencher as questões, acabaram por interromper o seu preenchimento e o retomaram bastante

---

<sup>1</sup> <<http://www.qualtrics.com/>>

tempo depois da interrupção. Contudo, não é possível determinar que os participantes mais lentos preencheram a pesquisa displicentemente, pois a pessoa também pode ter pesquisado sobre os assuntos relacionados ao conteúdo dos tuítes, por exemplo. Como não existe um indicador que invalide as respostas dos respondentes lentos e verificou-se que os respondentes muito rápidos possuem uma boa porcentagem de acerto na classificação dos tuítes em comparação com a amostra ambos os tipos de respondentes são mantidos na amostra.

Ao analisar os dados resolvi comparar o desempenho dos respondentes sob dois critérios, as respostas dadas pelos respondentes para as questões de controle e a autoavaliação dos respondentes a respeito de seus conhecimentos sobre segurança virtual.

Possui noção de segurança virtual?	Aprovado no controle	Reprovado no controle
Sim	68	9
Não	16	3

Tabela 3 – Autoavaliação dos respondentes da pesquisa sobre conhecimentos de ASVs x Real conhecimento de ASVs aferido pelo controle

Para entender melhor o perfil dos participantes da pesquisa eu extraí algumas informações da [Tabela 1](#) e as coloquei nas tabelas abaixo.

Respondente considera não ter noção de segurança virtual	Aprovado no controle	Reprovado no controle
	16 (84,21%)	3 (15,79%)

Tabela 4 – Divisão dos respondentes que afirmaram não ter noções de segurança entre dois grupos: os que passaram no controle e os que foram reprovados no controle

Possui noção de segurança virtual?	Reprovado no controle	
	Sim	Não
	9 (75,00%)	3 (25,00%)

Tabela 5 – Divisão dos respondentes que foram reprovados no controle entre dois grupos: os que afirmaram ter noções de segurança virtual e os que afirmaram não ter noções de segurança

A [Tabela 2](#) mostra que a maior parte (84,21%) dos respondentes que afirmaram não ter noções de segurança passaram no controle. Por outro lado, a [Tabela 3](#) mostra que 75% dos respondentes que foram reprovados no controle declararam ter noções de segurança virtual. Isso significa que a maior parte das pessoas que foram reprovadas no controle não estão cientes de que não possuem conhecimentos suficientes para identificar ASVs. No primeiro caso, o fato de que a maior parte das pessoas passaram no controle apesar de se considerarem com pouco conhecimento sobre segurança virtual pode significar que elas adotaram um critério bastante rigoroso pra avaliarem seus conhecimentos sobre o assunto.

	Porcentagem média de acerto na classificação dos tuítes
Tem noção de segurança	71,29% dos tuítes
Não tem noção de segurança	74,57% dos tuítes

Tabela 6 – Porcentagem média de acerto dos 10 tuítes classificados pelos respondentes dos grupos de respondentes auto declarados com ou sem noção de segurança virtual

	Porcentagem média de acerto na classificação dos tuítes
Aprovadas no controle	73,80% dos tuítes
Reprovadas no controle	55,00% dos tuítes

Tabela 7 – Porcentagem média de acerto dos 10 tuítes classificados pelos respondentes dos grupos de respondentes aprovados no controle e reprovados no controle

Como pode-se perceber nas tabelas 4 e 5, os grupos de respondentes possuem uma taxa de acerto de aproximadamente 70% na classificação dos tuítes, exceto o grupo das pessoas reprovadas no controle. Isso é esperado, pois os respondentes reprovados no controle possuem pouco conhecimento para identificar ASVs e por isso obtiveram um desempenho inferior aos demais grupos.

Acerto na classificação	Tuíte 3
Respondentes com noção de segurança	59 (76,62%)
Respondentes sem noção de segurança	13 (68,42%)

Tabela 8 – Número de pessoas que acertaram a classificação do Tuíte 3 para os grupos de respondentes que declararam ter conhecimentos de segurança virtual e dos que declararam não tê-los, acompanhado de sua respectiva porcentagem em relação ao seu respectivo grupo.

Acerto na classificação	Tuíte 3
Pessoas aprovadas no controle	65 (77,38%)
Pessoas reprovadas no controle	7 (58,33%)

Tabela 9 – Número de pessoas que acertaram a classificação do Tuíte 3 para os grupos de respondentes aprovados no controle e reprovados nele, acompanhado de sua respectiva porcentagem em relação ao seu respectivo grupo.

O Tuíte 3 é facilmente classificável, mas causa um pouco de dúvida o que se observa nas porcentagens de acerto dos grupos de respondentes mostradas pelas tabelas 6 e 7, pois se trata de um relatório (*report*) sobre ataques de negação de serviço (*DDoS*) em um determinado trimestre e não necessariamente uma ameaça ou ataque em curso a segurança virtual.

A taxa de acerto do Tuíte 4 é bastante alta em quase todos os grupos de respondentes, como mostram as tabelas 8 e 9, porque o conteúdo da mensagem trata de um tipo de ameaça bastante conhecida na área (*malwares*).

Acerto na classificação	Tuíte 4
Respondentes com noção de segurança	70 (90,90%)
Respondentes sem noção de segurança	18 (94,73%)

Tabela 10 – Número de pessoas que acertaram a classificação do Tuíte 4 para os grupos de respondentes que declararam ter conhecimentos de segurança virtual e dos que declararam não tê-los, acompanhado de sua respectiva porcentagem em relação ao seu respectivo grupo.

Acerto na classificação	Tuíte 4
Pessoas aprovadas no controle	79 (94,04%)
Pessoas reprovadas no controle	9 (75,00%)

Tabela 11 – Número de pessoas que acertaram a classificação do Tuíte 4 para os grupos de respondentes aprovados no controle e reprovados nele, acompanhado de sua respectiva porcentagem em relação ao seu respectivo grupo.

Acerto na classificação	Tuíte 1	Tuíte 5
Respondentes que afirmaram ter noções de segurança	33 (42,85%)	34 (44,15%)
Respondentes que afirmaram não ter noções de segurança	9 (48,57%)	8 (42,85%)

Tabela 12 – Número de pessoas, e sua respectiva porcentagem, que acertaram a classificação dos Tuítes 1 e 5 para o grupo dos respondentes que declararam ter conhecimentos de segurança virtual e dos que declararam não tê-los.

Acerto na classificação	Tuíte 1	Tuíte 5
Pessoas aprovadas no controle	38 (45,23%)	36 (42,85%)
Pessoas reprovadas no controle	4 (33,33%)	6 (50,00%)

Tabela 13 – Número de pessoas, e sua respectiva porcentagem, que acertaram a classificação dos Tuítes 1 e 5 para o grupo dos respondentes que foram aprovados no controle e dos que foram reprovados nele.

As tabelas 10 e 11 mostram que a taxa de acerto dos Tuítes 1 e 5 é baixa em todos os grupos de respondentes, o que é esperado. A dificuldade de identificar estes tuítes reside no fato de que alguns tipos de ASVs não são muito conhecidos, como os *zero-day exploits*, e pelo fato das notícias dos Tuítes 1 e 5 envolverem a Apple que é tida como uma empresa que produz softwares bastante seguros e por isso ASVs envolvendo produtos da empresa seriam pouco prováveis. O fato de os respondentes reprovados no controle terem maior porcentagem de acerto no Tuíte 5 em relação aos outros grupos de respondentes pode ser interpretado como uma anomalia resultante da dificuldade de se classificar o tuíte. Tal êxito pode ter sido obtido acidentalmente, dado que os reprovados no controle possuem menos conhecimentos sobre segurança virtual que os outros grupos participantes da pesquisa.

O tuíte número 10 pode ser erroneamente considerado como ASV, mas é um Spam.



A mensagem do tuíte possui uma notificação de atualização da base de dados de um antivírus (Avast). Porém, como o antivírus atualiza a base de dados automaticamente não é necessário que o usuário seja notificado da atualização via mensagem. Logo, a mensagem é interpretada como uma forma de se fazer propaganda do antivírus. O aviso só seria aceitável, e seria considerado um ASV, caso se tratasse de uma solução em particular para um problema gravíssimo e a empresa proprietária do software antivírus tivesse sido a primeira a encontrar a solução para o ASV em questão, por exemplo.

Alguns tuítes contêm expressões que possuem significado apenas em uma determinada cultura ou país. Por exemplo, a *Homeland Security* que seria algo como segurança do interior contra terrorismo nos Estados Unidos. Portanto, é necessário levar em conta que alguns participantes podem apresentar dificuldade para preencher a pesquisa devido a uma eventual baixa familiaridade com a língua inglesa ou à falta de compreensão de algumas expressões particulares dela ou de algum povo que a tem como sua língua pátria.

Tuítes mais fáceis e mais difíceis	1°	>	2°	>	3°	>	4°
Tuíte 3	AC		CN		SN		RC
Tuíte 4	SN		AC		CN		RC
Tuíte 1	SN		AC		CN		RC
Tuíte 5	RC		CN		AC		SN

Tabela 14 – Ordem decrescente das taxas de acerto dos 2 tuítes mais fáceis e os 2 mais difíceis obtidas por cada um dos grupos de respondentes. AC: Aprovado no Controle; RC: Reprovado no Controle; CN: Com noção de segurança virtual; SN: Sem noção de segurança virtual.

Nota de rodapé da Tabela acima<sup>2</sup>. Na monografia a nota de rodapé se comporta perfeitamente, mas no template que eu usei pra escrever este documento, não. Isso vai ser retirado quando esta análise for adicionada ao apêndice.

Pela tabela acima nota-se que o grupo AC classifica os tuítes fáceis melhor que os outros grupos, em média, e o grupo RC é o pior classificador para 3 dentre os 4 tuítes utilizados na comparação. Enquanto que o grupo CN alterna a 2° e 3° posições de melhor classificador para cada um dos tuítes utilizados na análise, o grupo SN alterna as posições de melhor e pior classificador.

Estes resultados corroboram a análise da [Tabela 1](#). Entre os respondentes que afirmaram não possuir conhecimentos de segurança, 75% deles não avaliaram adequadamente suas habilidades em identificar alertas de segurança virtual, pois este grupo de respondentes obteve, em média, um bom desempenho na classificação dos tuítes. O grupo de participantes que afirmaram ter um bons conhecimentos sobre segurança virtual obtiveram um desempenho mais estável em classificar os tuítes apesar de em média ser inferior

<sup>2</sup> AC foi considerado como maior que SN porque apesar de ambos terem apresentado a mesma porcentagem de acerto na classificação do Tuíte 5, o número de pessoas em AC que classificaram 5 corretamente é maior que o número de pessoas em SN.

ao desempenho dos respondentes que afirmaram não ter conhecimentos sobre segurança virtual. Os grupos aprovados e reprovados no controle corresponderam às expectativas de serem bons e ruins respectivamente em classificar os tuítes.

Desta forma, a análise dos resultados da pesquisa mostra que os respondentes não identificam bem o suficiente ASVs em língua inglesa e que a autoavaliação deles sobre o próprio conhecimento a respeito de segurança virtual é um pouco equivocada.

### A.3 Considerações finais

Como minha primeira vez a trabalhar com uma pesquisa piloto eu percebi que poderia tê-la feito um pouco diferente. Eu poderia montar um perfil técnico mais completo dos participantes que o realizado nesta pesquisa com a adição de perguntas que permitissem, por exemplo, inferir melhor o nível de conhecimento que os participantes da pesquisa possuem sobre segurança virtual.

Algumas das perguntas que ajudariam a montar um perfil técnico mais detalhado dos participantes da pesquisa poderiam incluir o tempo de experiência que a pessoa possui em cada função ou emprego em que a pessoa trabalhou com segurança. Outra informação interessante é o nível educacional do participante (superior completo, mestrado, etc). Também poderia ser pedido pra que eles definissem, mesmo que de forma simples o que eles entendem por ASV.